ARMY RESEARCH LABORATORY

Functional Estimation: The Asymptotic Regression Approach

by Joseph C. Collins III

ARL-TR-1644 March 1998

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5068

ARL-TR-1644 **March 1998**

Functional Estimation: The Asymptotic Regression Approach

A Dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics

Joseph C. Collins III
Survivability/Lethality Analysis Directorate, ARL

Abstract

Through an appeal to *asymptotic* Gaussian representations of certain empirical stochastic processes, we are able to apply the technique of continuous *regression* to derive parametric and nonparametric functional estimates for underlying probability laws.

This *asymptotic regression* approach yields estimates for a wide range of statistical problems, including estimation based on the empirical quantile function, Poisson process intensity estimation, parametric and nonparametric density estimation, and estimation for inverse problems.

Consistency and asymptotic distribution theory are established for the general parametric estimator. In the case of nonparametric estimation, we obtain rates of convergence for the density estimator in various norms.

We demonstrate the application of this methodology to inverse problems and compare the performance of the asymptotic regression estimator to other estimation schemes in a simulation study. The asymptotic regression estimates are easily computable and are seen to be competitive with other results in these areas.

Acknowledgments

I thank the following professionals at the U.S. Army Research Laboratory: Jerry Thomas, who gave me my first job at the Laboratory under the condition that I take a few statistics courses; Ballistic Vulnerability/Lethality Division Chief Paul Deitz, who has supported this program of study and research throughout the years; and Branch Chiefs Jill Smith, Lex Morrissey, Annie Young, and Richard Sandmeyer, who gave me free rein and allowed me to continue my studies. I am also indebted to Eric Edwards of LB&B Associates for his usual penetrating and thorough editorial review.

Furthermore, I wish to express sincere appreciation for the University of Delaware faculty members of my Dissertation Committee—Paul Eggermont, Vincent LaRiccia, David Mason, and Zuhair Nashed—all of whom have given excellent instruction and guidance both in and out of the classroom.

Vince, acting as my advisor and Dissertation Committee chair, has provided countless hours of sound counsel in technical, professional, and personal matters and always made sure that I not lose sight of the goal and purpose of study and of life in general. I regard him, with honor, as mentor and friend.

Finally, I thank my children and my parents alike, who, without understanding, have trusted me when I told them that I really would be finished with school some day. ॐ

INTENTIONALLY LEFT BLANK.

Table of Contents

Page

List of Figures

Intentionally Left Blank.

List of Tables

INTENTIONALLY LEFT BLANK.

1. Asymptotic Regression Estimation

1.1 Background

Arguably, *estimation* is the essential problem of statistics. To set up the conceptual framework for estimation, we assume that a random quantity T behaves according to a certain but unknown probability law $\mathcal{L}$ and that quantitative information about $\mathcal{L}$ can be recovered through observation of T. In the English language, the noun *estimate* connotes a subjective judgment that takes the place of definite factual knowledge. This is akin to its use in statistics, which we illustrate with the typical example of an independent, identically distributed (i.i.d.) random sample.

> The random variable T has a probability density function given by $f(\cdot;\tau)$ where the parameter τ is unknown. We observe a random sample $(T_1,\ldots,T_n)$ of i.i.d. values of T. A specific function $\tau_n(T_1,\ldots,T_n)$ of the observation may be called an *estimator* of τ. Given an observation, the value of that function is then an *estimate* of τ. The estimate is considered to be an acceptable substitute for the unknown true parameter value.

Efforts to quantify estimator accuracy and remove, reduce, or otherwise control the subjective element in this process can be traced back at least two hundred years. Since that time, numerous criteria for generating and selecting estimators have been proposed, including the methods of least squares (due to K. F. Gauss [20]), moments (due to K. Pearson [55]), maximum likelihood (due to R. A. Fisher [19]), minimum chi-square, minimum distance, maximum product of spacings, minimax, Bayes, and Pitman, to name a few.

In spite of this diversity, however, statisticians seem to agree that the accuracy of a useful estimator should increase as the sample size n increases. This requires among other things that $f(\cdot;\tau) \mapsto \tau$ is a function, which is to say that two distinct parameter values cannot correspond to the same distribution. This condition, termed *identifiability*, is a feature of a class of probability distributions and not of any estimator.

Given an identifiable class of distributions, we can begin to talk about desirable properties of estimators. First of all, note that an estimator is calculated from a finite sample that is representative of a larger finite or possibly infinite population. Fisher's [19] original characterization of the notion of *consistency* is that an estimator calculated from the entire population should achieve the true parameter value. Of course, the estimator itself is also a random variable (r.v.); and, formally, an estimator is considered to be consistent if it *converges in probability* to the true parameter value as the sample size increases. That is,

$$(\forall \varepsilon > 0)\,(\forall \delta > 0)\,(\exists N)\,(\forall n > N) \quad \Pr(|\tau_n - \tau| > \delta) < \varepsilon,$$

which is denoted more succinctly by

$$\tau_n \xrightarrow{p} \tau \;\text{ as }\; n \to \infty.$$

A concept related to convergence in probability is that an estimator may be *asymptotically unbiased.* That is,

$$\mathrm{E}\,\tau_n \to \tau \;\text{ as }\; n \to \infty,$$

where "E" denotes the expectation. For an asymptotically unbiased estimator, the criterion that accuracy increases with sample size can be formalized as

$$\mathrm{Var}\,\tau_n \to 0 \;\text{ as }\; n \to \infty,$$

where "Var" denotes the variance. This states that the dispersion of an estimator should decrease as n increases. The connection between accuracy and variance of an estimator was recognized in the time of Gauss and Laplace, as noted by Stuart and Ord [72].

The existence of the well-known Cramér-Rao lower bound for the variance of an estimator provides in certain cases a criterion for optimality of an estimator. This occurs, for example, in maximum-likelihood estimation of a probability density function parameter based on a random sample. See

C. R. Rao [57] for an exposition. The maximum-likelihood estimator (MLE) achieves this bound asymptotically, and is hence asymptotically optimal. As a special case, we have the problem of regression in the linear model with normal error,

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where Y is an n-vector, X is a fixed $n \times p$ matrix, β is a p-vector of coefficients to be estimated, and the error vector ε has the multivariate normal distribution with $\varepsilon \sim N_n(0, \sigma^2 I)$. In this case, it is known that the maximum-likelihood, least-squares, and minimum-variance unbiased estimators of β are one and the same. Seber [60] points this out.

In the present work, we propose an estimation methodology for a class of problems that includes estimation of τ based on i.i.d. observations $T_1, \ldots, T_n$ from the probability density $f(\cdot; \tau)$. The scheme is based on a generalization of the linear regression model of equation (1.1).

1.2 Introduction

The estimation procedure proposed in this work is based on the observation of a stochastic process with certain asymptotic properties. The principle of maximum likelihood and the technique of continuous-time regression are applied to an asymptotic version of the observed process to yield estimates of an underlying probability law. The resulting estimators have optimal properties similar to those of maximum-likelihood and least-squares estimators.

We can now describe the modeling situation and estimation procedure that are the basis of this work. To that end, let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with common probability law $\mathcal{L}(\tau)$, where the unknown true parameter value τ lies in some suitable parameter space Θ. It is τ that we wish to estimate. For each n, let $X_n(t)$ be a stochastic process with sample paths in a space $\mathcal{S}$ of functions defined on a domain I. Suppose that X_n is determined by $(T_1, \ldots, T_n)$ and that X_n is a sufficient statistic for $\mathcal{L}(\tau)$. Furthermore, let the sequence of stochastic processes $\{X_n\}_{n\in\mathbb{N}}$ converge in

distribution in the sense that

$$\sqrt{n}(X_n - M_\tau) \xrightarrow{d} A_\tau \quad \text{as} \quad n \to \infty, \tag{1.2}$$

where M_τ is a deterministic function and A_τ is a zero-mean Gaussian stochastic process with covariance function $\mathrm{E}\, A_\tau(s)A_\tau(t) = K_\tau(s,t)$.

For a finite-dimensional (vector) random variable, convergence in distribution is taken to mean convergence of the cumulative distribution at each continuity point. The infinite-dimensional (function) situation is more complicated, and the practical technical details are application-dependent. Our discussion of convergence in distribution for random functions is deferred to section 2.6.1.

The following model, which we call the asymptotic model for the process X_n, plays a central role in the development of the proposed estimator. With M_τ and A_τ as in (1.2), consider a process X_n^* defined by

$$X_n^*(t) = M_\tau(t) + \tfrac{1}{\sqrt{n}} A_\tau(t). \tag{1.3}$$

The key feature of this model is that the mean and covariance functions of the process sequence elements share a common parameter τ.

In the remainder of this chapter, we outline a general estimation technique for the unknown parameter of the asymptotic model (1.3) based on concepts from continuous-time regression and maximum-likelihood estimation for Gaussian stochastic processes. The technique is applicable in the finite-dimensional case, where $\Theta \subseteq \mathbb{R}^d$ for some finite positive integer d, and also in the nonparametric setting, where Θ is some space of functions on I. The proposed estimator for τ based on the X_n of (1.2) is then obtained by using X_n in place of X_n^* in the estimation scheme. (In what follows, we only occasionally distinguish X_n from X_n^*.) The properties of these estimators are studied in the remainder of this work.

In chapter 2, we consider the parametric estimation problem, in which the parameter space is a subset of $\mathbb{R}^d$ for a finite positive integer d. We also discuss the existence, consistency, and optimality of our estimators. In chapter 3, we consider nonparametric estimation and see that the corresponding

penalized estimation procedure yields a solution with desirable properties and provides a unified approach to a wide class of statistical problems. In chapter 4, we discuss the practical computational aspects of the proposed nonparametric estimation scheme.

Fundamental examples of stochastic processes that satisfy (1.2) are the empirical cumulative distribution function (c.d.f.), empirical quantile function (q.f.), and Poisson counting process, as detailed in the following examples.

Example (Random Sample). Let $T_1, \ldots, T_n$ be i.i.d. random variables defined on I with continuous cumulative distribution function $F = F_\tau$ and probability density function $f = F'$. The empirical c.d.f. is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \leqslant t).$$

It is well known that F_n is a sufficient statistic for τ and that

$$\sqrt{n}\,(F_n - F) \xrightarrow{d} B \circ F \quad \text{as} \quad n \to \infty.$$

Here, B is a standard Brownian bridge, which is a zero-mean Gaussian process with covariance function $s \wedge t - st$. See, for example, Billingsley [5], Theorem 16.4. So we identify $F_n(t)$ with $F(t) + n^{-1/2}\, B(F(t))$, and then we estimate τ based on modeling F_n by equation (1.3) with the indicated mean and covariance functions. Specifically, the model is

$$F_n(t) = F_\tau(t) + \tfrac{1}{\sqrt{n}} A_\tau(t),$$

where A_τ is a zero-mean Gaussian process with covariance function $F_\tau(s \wedge t) - F_\tau(s)F_\tau(t)$.

Nussbaum [48] exhibits the following result, which provides theoretical justification for the asymptotic identification of an empirical process with a Gaussian model in the case of density estimation: The two sequences of statistical experiments given by observations

$$T_i,\ i \in \{1, \ldots, n\},\ \text{i.i.d. with p.d.f. } f, \quad \text{and}$$
$$dX(t) = f^{1/2}(t)\, dt + \tfrac{1}{2} n^{-1/2}\, dW(t),$$

where W is a standard Wiener process, are asymptotically equivalent in the sense of Le Cam's deficiency distance. Le Cam and Yang [43] discuss the asymptotic characterization of statistical experiments. Our aim, however, is to exploit the general heuristic identification of (1.2) with (1.3) and thereby obtain practical and useful estimation procedures for a variety of statistical models. And so we continue.

For the random sample, a second model is based on another sufficient statistic, the empirical quantile function. The quantile function Q, defined by $Q(u) = \inf\{t : F(t) \geqslant u\}$, is the unique left-continuous pseudo-inverse of F. The empirical quantile function is given by

$$Q_n(u) = \inf\{t : F_n(t) \geqslant u\},$$

and the density quantile function is $g = f \circ Q$. Differentiation of $F(Q(u)) = u$ yields $Q' = 1/g$. In this case, it is known that

$$\sqrt{n}\frac{1}{Q'}(Q_n - Q) \xrightarrow{d} B \quad \text{as} \quad n \to \infty,$$

where B is a standard Brownian bridge. See, for example, Shorack and Wellner [63], section 18.1. So we identify $Q_n(t)$ with $Q_\tau(t) + n^{-1/2} Q'_\tau(t) B(t)$ and use the model

$$Q_n(t) = Q_\tau(t) + \tfrac{1}{\sqrt{n}} A_\tau(t),$$

where A_τ is a zero-mean Gaussian process with covariance function $Q'_\tau(s)Q'_\tau(t)(s \wedge t - st)$.

Example (Poisson Process). Let $T_1, \ldots, T_n$ be i.i.d. Poisson processes on $I = [0, 1]$. Each T_i has a representation as a sum of point masses

$$T_i = \sum_{j=1}^{k_i} \delta_{t_{ij}},$$

and with each T_i we associate the counting process

$$N_i(t) = \sum_{j=1}^{k_i} I(t_{ij} \leqslant t).$$

The processes N_i have a common *compensator*, or mean-value function, $G(t) = G_\tau(t) = \mathrm{E}\, N_i(t)$. Its derivative $g = G'$ is called the *intensity* function. A sufficient statistic for τ is

$$X_n(t) = \frac{1}{n} \sum_{i=1}^{n} N_i(t).$$

It is known that

$$\sqrt{n}\,(X_n - G) \xrightarrow{d} \sqrt{G(1)} \cdot W \circ \frac{G}{G(1)} \quad \text{as} \quad n \to \infty,$$

where W is a standard Wiener process, which is a zero-mean Gaussian process with covariance function $\mathrm{E}\, W(s)W(t) = s \wedge t$. So, we identify $X_n(t)$ with $G(t) + G(1)^{1/2} n^{-1/2} W\,[G(t)/G(1)]$ and model the process as

$$X_n(t) = G(t) + \tfrac{1}{\sqrt{n}} A(t),$$

where A is a zero-mean Gaussian process with covariance function $G(s \wedge t)$.

Other examples of applications that may fit into this scheme include estimation based on the hazard function, random censoring models, marked Poisson processes, and deconvolution and other ill-posed inverse problems.

A special case of this procedure is noted by Emanuel Parzen [54]. He considers problems of location and scale estimation based on continuous-time regression of the empirical quantile function. His methodology reproduces well-known results about the use of linear combinations of order statistics to solve such problems. For distributions with location and scale dependence, such as the three-parameter lognormal and Weibull distributions, Kindermann and LaRiccia [32] propose an easily computable generalization of Parzen's procedure.

As stated, the estimation technique we propose is based on ideas from continuous-time regression. So, before the estimators are defined, we give a brief overview of that subject. We also develop the generalizations that enable us to apply continuous-time regression to the modeling of this section's examples.

1.3 Regression for Gaussian Processes with Known Covariance

This section contains the basic facts we need about continuous-time regression for Gaussian stochastic processes with known covariance functions. More details are presented in the Appendix. The development is adapted from the work of Emanuel Parzen [51], [52], and [53].

Consider a Gaussian stochastic process

$$X(t) = M(t) + A(t)$$

defined for $t \in I$ with unknown mean-value function $\mathrm{E}\, X(t) = M(t)$ and known covariance function $\mathrm{E}\, A(s)A(t) = K(s,t)$. The purpose of continuous-time regression is estimation of the mean-value function. This is accomplished by identifying an appropriate likelihood ratio and then conducting maximum-likelihood estimation.

The reference measure for the likelihood ratio is derived from another process $Y(t)$, which is a zero-mean Gaussian stochastic process with covariance function $\mathrm{E}\, Y(s)Y(t) = K(s,t)$, so X and Y have the same (known) covariance function. Denote by $\mathcal{P}(K,M)$ and $\mathcal{P}(K)$ the probability measures induced by $X(t)$ and $Y(t)$, respectively, on the space of sample paths. The likelihood ratio itself is the Radon-Nikodym derivative

$$\frac{d\mathcal{P}(K,M)}{d\mathcal{P}(K)},$$

which is defined using a space of random variables $L_2(X)$, a function space H_K, and a map ϕ between the two.

To define $L_2(X)$, first consider the linear span of $X(t)$, denoted $L(X)$ and defined by

$$L(X) = \left\{ \sum_{i=1}^{n} a_i X(t_i) : n \in \mathbb{N},\ t_i \in I,\ a_i \in \mathbb{R} \right\}.$$

This is the set of all finite linear combinations of values of X taken at ar-

bitrary points. An inner product on $L(X)$ is given by $\langle u, v \rangle = Euv$. The Hilbert space $L_2(X)$ is the completion of $L(X)$ in the corresponding norm $\|u\| = \sqrt{Eu^2}$.

Denote by H_K the reproducing kernel Hilbert space (RKHS) of functions on I with reproducing kernel K, inner product $\langle \cdot, \cdot \rangle_K$, and norm $\|\cdot\|_K$, where, of course, $\|x\|_K^2 = \langle x, x \rangle_K$. See Aronszajn [2] for a discussion of reproducing kernel Hilbert spaces. The point evaluation functional representers in H_K are denoted by K_t, $t \in I$, where $K_t(s) = K(s,t)$. These have the "reproducing" property—namely, $\langle K_t, f \rangle_K = f(t)$ for all $t \in I$ and $f \in H_K$, and in particular $\langle K_s, K_t \rangle = K(s,t)$ for all s and t in I.

The function $\phi_K : H_K \to L_2(X)$ is defined on the generators of H_K by $\phi_K(K_t) = X(t)$ and by linear extension on the whole space. The map ϕ is in fact a congruence, or inner-product-preserving vector space isomorphism.

We are now able to define the likelihood ratio.

Theorem 1.1. *In the event that $M \in H_K$, the measures $\mathcal{P}(K, M)$ and $\mathcal{P}(K)$ are equivalent. Then their likelihood ratio is given by*

$$L(M) = \frac{d\mathcal{P}(K,M)}{d\mathcal{P}(K)}(X) = \exp\left[\phi_K(M) - \tfrac{1}{2}\|M\|_K^2\right]. \tag{1.4}$$

Proof. See Parzen [51]. □

This functional is the basis for determining maximum-likelihood estimates of the parameter M. Specifically, one takes as the estimator any value $\hat{M}$ that is a solution of the optimization problem

$$\underset{M}{\text{maximize}}\ L(M) \ \text{ subject to } \ M \in H_K.$$

Thus, the estimator $\hat{M}$ satisfies

$$\frac{d\mathcal{P}(K,\hat{M})}{d\mathcal{P}(K)}(X) = \sup\left\{\frac{d\mathcal{P}(K,M)}{d\mathcal{P}(K)}(X) : M \in H_K\right\},$$

which is typically an ill-posed problem. To avoid this difficulty, one usually specifies a set $\mathcal{M}$ of candidate functions for M and then attempts to solve a problem equivalent to

$$\underset{M}{\text{maximize}}\ L(M) \ \text{ subject to } \ M \in \mathcal{M}. \tag{1.5}$$

In this case, the estimator $\hat{M}$ satisfies

$$\frac{d\mathcal{P}(K,\hat{M})}{d\mathcal{P}(K)}(X) = \sup\left\{\frac{d\mathcal{P}(K,M)}{d\mathcal{P}(K)}(X) : M \in \mathcal{M}\right\}.$$

It is tempting to observe that in a formal sense $\phi_K(K_t) = \langle K_t, X\rangle_K = X(t)$, and so $\phi_K(f) = \langle f, X\rangle_K$ for all $f \in H_K$ as another expression of the reproducing property of the K_t. Then one would have

$$\begin{aligned} -2\log L(M) &= -2\phi_K(M) + \|M\|_K^2 \\ &= -2\langle X, M\rangle_K + \|M\|_K^2 \\ &= \|X - M\|_K^2 - \|X\|_K^2, \end{aligned} \tag{1.6}$$

and the optimization problem (1.5) would be a true least-squares problem. However, the sample paths of X do not necessarily lie in H_K, the construction $\phi_K(f)$ cannot be an inner product of elements in H_K, and the characterization of (1.6) is only formal. Some authors use the inner product notation for ϕ, with the caveat that it is not really an inner product. We reserve inner product notation for inner products and denote the congruence by ϕ. Nonetheless, in all of our applications, the congruence does indeed have the same form as the inner product.

We now consider several standard parameter set configurations, or possibilities for the set $\mathcal{M}$. The first two, parametric and nonparametric estimation, are the ones used in the current work. The other three are included for their intrinsic interest and to illustrate the connection between continuous regression for Gaussian processes and discrete finite least-squares regression. The standard parameter set configurations follow.

(i) Nonparametric Estimation. $\mathcal{M}$ is a subset of some function space such as L_1. This is the most general case, in terms of the restrictions placed on candidate functions.

(ii) Parametric Estimation. In this case $\mathcal{M} = \{M_\tau : \tau \in \Theta\}$ for some (finite-dimensional) set $\Theta \subseteq \mathbb{R}^d$ and family of parametric functions $M_\tau(t)$. The estimator M_n usually exists, as long as $\mathcal{M}$ is a reasonable parametric family. But calculation can be troublesome, and determining the probabilistic properties of the estimator in complete generality can be practically impossible. This is not so in the following situation, which is a special case of *(ii)*.

(iii) Finite-Dimensional Subspace. For a fixed positive finite integer k, choose the functions $f_i, \ldots, f_k$ in H_K and let $\mathcal{M} = \left\{\sum_{i=1}^k a_i f_i : a_i \in \mathbb{R}\right\}$. Then one can show that the estimator is given by

$$M_n(t) = \sum_{i=1}^{k} A_i f_i(t),$$

in which the vector $A = (A_1, \ldots, A_k)^T$ is a solution of the normal equations $CA = B$, where the matrix C and vector B have components $C_{ij} = \langle f_i, f_j \rangle_K$ and $B_i = \phi_K(f_i)$, respectively. Note the similarity to linear regression. In this case, M_n is a uniformly minimum-variance unbiased estimator. See Parzen's papers for the details.

Comments on the utility of the following two examples range from "illuminating" in Parzen [51], section 8.36, to "of little interest" in Grenander [22], p. 98.

(iv) Finite Domain. Consider $X(t) = M(t) + A(t)$ with $t \in \{t_1, \ldots, t_n\}$. Then X, M, and A are finite-dimensional (column) vectors, so we can write them in terms of their components; i.e., $X = (X_1, \ldots, X_n)^T$, where $X_i = X(t_i)$, and likewise for M and A. Thus, the model becomes

$$X = M + A,$$

with $\mathrm{E}\, X = M$ and $A \sim N_n(0, K)$. The vector A has the multivariate normal distribution with variance-covariance matrix $K = \mathrm{E}\, AA^T$, where the compo-

nents of K are $K_{ij} = \mathrm{E}\, A_i A_j$. In this case, $H_K = \mathbb{R}^n$ with the interpretation that $f_i = f(t_i)$ for $f = (f_1, \ldots, f_n)^T \in H_K$. The inner product in H_K is given by $\langle f, g \rangle_K = f^T K^{-1} g$. The point evaluation functional representers in H_K are the columns of K. Specifically, let K_i be the i^{th} column of K. Then the requisite properties $\langle K_i, K_j \rangle_K = K_{ij}$ and $\langle f, K_i \rangle_K = f_i$ are satisfied. The congruence ϕ_K is given by $\phi_K(f) = \langle f, X \rangle_K$, so that $\phi_K(K_i) = X_i$, as required. The likelihood ratio for X with respect to a zero-mean multivariate normal random variable having the same variance-covariance matrix is simply the quotient of the appropriate multivariate normal probability densities,

$$L = \frac{(2\pi|K|)^{-1/2} \exp[-\frac{1}{2}(X-M)^T K^{-1}(X-M)]}{(2\pi|K|)^{-1/2} \exp[-\frac{1}{2} X^T K^{-1} X]}.$$

It is easily verified that $L = \exp\left[\phi_K(M) - \frac{1}{2}\|M\|_K^2\right]$. Thus, the continuous-regression setup reduces to the familiar maximum-likelihood formulation for finite regression, and the estimator $\hat{M}$ is given by

$$\hat{M} = \arg\min_{M \in \mathcal{M}} \|X - M\|_K^2.$$

Note that this is a true least-squares problem. A specific form for $\mathcal{M}$ is considered in the final example.

(v) Linear Model. In case *(iv)*, fix an $n \times k$ matrix Z and let

$$\mathcal{M} = \{Z\beta : \beta \in \mathbb{R}^k\}.$$

Then it is very well known that the estimator

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^k} \|X - Z\beta\|_K^2$$

is any solution β of the normal equations

$$Z^T K^{-1} Z \beta = Z^T K^{-1} X.$$

In the nonsingular case, the unique solution is

$$\hat{\beta} = (Z^T K^{-1} Z)^{-1} Z^T K^{-1} X.$$

Of course, this is the classical linear regression model.

In our applications, we only consider the parameter set configurations of cases *(i)* and *(ii)*. The remaining cases simply illustrate the connection between continuous and discrete regression for Gaussian processes and are included to illustrate the fact that continuous regression is a generalization of the familiar discrete (finite) situation.

1.4 Sequences of Processes with Known Covariance

The basic model for continuous regression, described in section 1.3, is

$$X(t) = M(t) + A(t)$$

with $\mathrm{E}\,X(t) = M(t)$ unknown and $\mathrm{E}\,A(s)A(t) = K(s,t)$ known. Two generalizations of this model are required in order to apply the methodology of continuous regression to the models described in section 1.2. In this section, we discuss the first required generalization; that is, we identify the likelihood ratio sequence for the known-covariance version of the model sequence (1.3). The second generalization, adaptation of the continuous-regression methodology to the case of unknown covariance, is discussed in section 1.5. Finally, in section 1.6, we combine the two generalizations and formulate the estimation principle that is the subject of this work.

Now we compute the likelihood ratio sequence

$$\frac{d\mathcal{P}(K_n, M)}{d\mathcal{P}(K)}(X_n)$$

for the sequence of models

$$X_n(t) = M(t) + \tfrac{1}{\sqrt{n}}A(t). \tag{1.7}$$

The mean-value function here is $\mathrm{E}\,X_n(t) = M(t)$, and the (known) covariance functions K_n satisfy

$$K_n(s,t) \equiv \mathrm{Cov}[X_n(s), X_n(t)] = \tfrac{1}{n}K(s,t).$$

As in section 1.3, $\mathcal{P}(K_n, M)$ and $\mathcal{P}(K)$ are the probability measures induced by X_n and Y, respectively, on the space of sample paths. Here, $Y(t)$ is a zero-mean Gaussian process with covariance function $\mathrm{E}\,Y(s)Y(t) = K(s,t)$.

First, observe that $nK_n = K$. It is clear that $H_K = H_{K_n}$. The inner product of H_K is characterized by $\|K_t\|_K^2 = K(t,t)$, whereas the inner product on H_{K_n} is characterized by $\|K_{nt}\|_{K_n}^2 = K_n(t,t)$. Then we have

$$\|K_t\|_K^2 = K(t,t) = nK_n(t,t) = n\|K_{nt}\|_{K_n}^2 = n\left\|\tfrac{1}{n}K_t\right\|_{K_n}^2 = \tfrac{1}{n}\|K_t\|_{K_n}^2,$$

and so for all $G \in H_K$

$$\|G\|_{K_n}^2 = n\|G\|_K^2.$$

Next, we consider the sequence of maps $\phi_n : H_{K_n} \to L_2(X_n)$, which satisfy

$$\phi_n(K_{nt}) = X_n(t)$$

for all $n \in \mathbb{N}$, where $K_{nt} = K_n(\cdot, t)$. We make the dependence of ϕ_K upon X explicit and recall that $\phi_K : H_K \to L_2(X)$ is characterized by $\phi_K(K_t) = \phi_K(X, K_t) = X(t)$. The map $\phi_n : H_K \to L_2(X_n)$ likewise satisfies $\phi_n(K_{nt}) = \phi_n(X_n, K_{nt}) = X_n(t)$, so by linearity we get $\phi_n(X_n, K_t) = \phi_n(X_n, nK_{nt}) = nX_n(t)$. Assuming the formal dependence of the maps on the processes is independent of n, which is true in all practical situations, we have

$$\phi_n(X_n, G) = n\phi_K(X_n, G).$$

Then the likelihood ratio sequence for (1.7) is given by

$$\begin{aligned}\frac{d\mathcal{P}(K_n, M)}{d\mathcal{P}(K)}(X_n) &= \exp\left[n\phi_K(X_n, M) - \tfrac{n}{2}\|M\|_K^2\right] \\ &= \left[\frac{d\mathcal{P}(K, M)}{d\mathcal{P}(K)}(X_n)\right]^n. \qquad (1.8)\end{aligned}$$

Of course, a maximum-likelihood estimator $\hat{M}$ of M satisfies

$$\frac{d\mathcal{P}(K_n, \hat{M})}{d\mathcal{P}(K)}(X_n) = \sup\left\{\frac{d\mathcal{P}(K_n, M)}{d\mathcal{P}(K)}(X_n) : M \in \mathcal{M}\right\}.$$

Therefore, in light of the scaling property of equation (1.8), we finally observe that a maximum-likelihood estimator $\hat{M}$ of M also satisfies

$$\frac{d\mathcal{P}(K,\hat{M})}{d\mathcal{P}(K)}(X_n) = \sup\left\{\frac{d\mathcal{P}(K,M)}{d\mathcal{P}(K)}(X_n) : M \in \mathcal{M}\right\}$$

for suitable $\mathcal{M}$. Thus, for the purposes of optimization, the dependence of the likelihood ratio on n may be ignored. We simply use the likelihood ratio (1.4).

1.5 Regression for Gaussian Processes with Unknown Covariance

The results of sections 1.3 and 1.4 pertain to Gaussian processes with known covariance. They are not directly applicable to the modeling situations of section 1.2. In this section, we propose an iterative estimation scheme for a certain class of Gaussian processes (with unknown covariance) that does include the models of section 1.2.

Consider a Gaussian stochastic process

$$X(t) = M(t) + A(t)$$

with unknown mean $\mathrm{E}\,X(t) = M(t)$ and covariance $\mathrm{E}\,A(s)A(t) = K_M(s,t)$. Note that the covariance depends on the unknown mean. Equivalently, the mean and covariance functions share a common unknown parameter that we wish to estimate, as in the models of section 1.2. We assume that the true mean-value function M lies in some fixed set $\mathcal{M}$ of candidate mean-value functions.

We construct a recursive sequence $(M_0, M_1, M_2, \ldots)$ of estimators for M as follows. Select an arbitrary $M_0 \in \mathcal{M}$, and for $i \geqslant 1$, let $M_i \in \mathcal{M}$ be such that

$$\frac{d\mathcal{P}(K_{M_{i-1}},M_i)}{d\mathcal{P}(K_{M_{i-1}})}(X) = \sup\left\{\frac{d\mathcal{P}(K_{M_{i-1}},M)}{d\mathcal{P}(K_{M_{i-1}})}(X) : M \in \mathcal{M}\right\}.$$

In words, we first assume an "initial guess" parameter value M_0. Then we repeatedly calculate the covariance function using the current parameter value

and estimate a new parameter value by means of the principle of maximum likelihood for Gaussian processes with known covariance described in section 1.3.

1.6 The Asymptotic Regression Estimation Principle

We can now formulate the general principle for estimation of the unknown parameter of section 1.2.

> Apply the recursive estimation scheme of section 1.5 to the probability model sequence given by equation (1.3), taking into account the scaling property of section 1.4.

Explicitly, the asymptotic model sequence is

$$X_n^*(t) = M(t) + \tfrac{1}{\sqrt{n}} A(t),$$

where $A(t)$ is a zero-mean Gaussian process with covariance function $\mathrm{E}\, A(s)A(t) = K_M(s,t)$. The covariance function depends on the unknown mean-value function. Equivalently, the mean and covariance functions share a common parameter. Our aim is to estimate this unknown parameter.

We assume that $\mathcal{M}$ is a fixed set of candidate mean-value functions and that the process sequence $\{X_n\}_{n\in\mathbb{N}}$ is given. The definition follows.

Definition (Asymptotic Regression Estimator). For each n, construct the recursive sequence $(M_{n,0}, M_{n,1}, M_{n,2}, \dots)$ of estimators for M in this manner: Select an arbitrary $M_{n,0} \in \mathcal{M}$, and for $i \geqslant 1$, let $M_{n,i} \in \mathcal{M}$ be such that

$$\frac{d\mathcal{P}(K_{M_{n,i-1}}, M_{n,i})}{d\mathcal{P}(K_{M_{n,i-1}})}(X_n) = \sup\left\{ \frac{d\mathcal{P}(K_{M_{n,i-1}}, M)}{d\mathcal{P}(K_{M_{n,i-1}})}(X_n) : M \in \mathcal{M} \right\}.$$

We use the terms *asymptotic regression estimator*, *AR estimator*, and *ARE* to refer to any element of a sequence so obtained.

We intend to show that for fixed arbitrary $i \geqslant 1$, the AR estimator has good properties as $n \to \infty$. Carroll and Ruppert [7] have proposed a

similar iterative estimation scheme for the problem of nonlinear regression with heteroscedastic error. We hope to show for our problem, as they have done for theirs, that stopping the procedure after a small number of steps results in an estimator with reasonable small-sample properties.

In chapter 2, we consider the parametric estimation problem, in which the parameter space Θ is a subset of $\mathbb{R}^d$ for some finite positive integer d. We discuss the existence, consistency, and large-sample distributions of AR estimators. Results in this chapter establish the asymptotic optimality of AR estimators.

In chapter 3, we consider nonparametric AR estimation. We see that the solution of the corresponding penalized problem is an estimator with desirable properties. Also, we see that AR estimation provides a unified approach to a wide class of statistical problems.

In chapter 4, we discuss the practical computational aspects of nonparametric AR estimation. Topics here include discretization, software implementation, and a reliable data-driven method for smoothing parameter selection.

INTENTIONALLY LEFT BLANK.

2. Parametric AR Estimation

2.1 General Parametric Estimation

In this chapter, we consider asymptotic regression estimation in the case of a finite-dimensional real parameter space. In this section, we define the parametric AR estimator and its associated optimization problem. In section 2.2, we present and discuss the main results concerning the consistency and asymptotic properties of the AR estimator. Sections 2.3, 2.4, and 2.5 describe the application of AR estimation to some standard statistical modeling situations. Section 2.6 contains technical material including the proofs of the theorems in section 2.2.

To cast the AR model and estimation procedure of section 1.2 into the parametric setting, we suppose, as always, that X_n is a sequence of stochastic processes with sample paths in a space $\mathcal{S}$ of functions defined on a domain I. In this chapter, we denote the true parameter value by τ and assume that $\tau \in \Theta \subseteq \mathbb{R}^d$ where d is a positive integer.

Now let $A = A_\tau$ be a Gaussian process with mean value $\mathrm{E}\, A = 0$ and continuous positive-definite covariance function $\mathrm{E}\, A(s)A(t) = K_\tau(s,t)$, and let M_τ be a deterministic function on I. Next, define $A_{n,\tau} = \sqrt{n}(X_n - M_\tau)$ and suppose that

$$A_{n,\tau} \xrightarrow{d} A_\tau \quad \text{as} \quad n \to \infty. \tag{2.1}$$

In order to define the AR estimator, we need to identify certain spaces, operators, and norms. So for any $\theta \in \Theta$, let H_θ be the RKHS with reproducing kernel K_θ, inner product $\langle \cdot, \cdot \rangle_\theta$, norm $\| \cdot \|_\theta$, and point evaluation functional representers $K_{\theta t}$. We require that, while the norms may depend on θ, the underlying spaces remain constant. That is to say, $H_\theta = H_\tau = H$ for all $\theta \in \Theta$. Furthermore, suppose that $M_\theta \in H$ for each $\theta \in \Theta$. Let the bilinear functional $\phi_\theta : \mathcal{S} \times H \to \mathbb{R}$ satisfy

$$\phi_\theta(Z, K_{\theta t}) = Z(t)$$

for each $t \in I$. Since $\phi_\theta(f,g) = \langle f,g\rangle_\theta$ if $f \in H_\theta$, this map is given by the inner product when both of its arguments lie in H_θ.

Let γ be a fixed element of Θ. The probability density functional for a Gaussian stochastic process Z with mean M_θ and covariance K_γ with respect to a mean-zero Gaussian process having the same covariance is

$$\frac{d\mathcal{P}(\gamma,\theta)}{d\mathcal{P}(\gamma)}(Z) = \exp\left[\phi_\gamma(Z,M_\theta) - \tfrac{1}{2}\|M_\theta\|_\gamma^2\right]. \tag{2.2}$$

With $n \in \mathbb{N}$ and X_n both fixed, we take as an estimator of τ any $\tau_n \in \Theta$ satisfying

$$\frac{d\mathcal{P}(\gamma,\tau_n)}{d\mathcal{P}(\gamma)}(X_n) = \sup\left\{\frac{d\mathcal{P}(\gamma,\theta)}{d\mathcal{P}(\gamma)}(X_n) : \theta \in \Theta\right\}.$$

We now define the AR estimator sequence.

Definition (Parametric AR Estimator). For fixed n, X_n, and $\tau_{n,0} \in \Theta$, the recursive sequence of AR estimators $(\tau_{n,0}, \tau_{n,1}, \tau_{n,2}, \dots)$ is defined for all $i \in \mathbb{N}$ by

$$\frac{d\mathcal{P}(\tau_{n,i-1},\tau_{n,i})}{d\mathcal{P}(\tau_{n,i-1})}(X_n) = \sup\left\{\frac{d\mathcal{P}(\tau_{n,i-1},\theta)}{d\mathcal{P}(\tau_{n,i-1})}(X_n) : \theta \in \Theta\right\}.$$

We occasionally write this as $\tau_{n,i} = S_n(\tau_{n,i-1})$ for notational convenience.

Maximization of the likelihood ratio (2.2) is equivalent to minimization of

$$J_{n,\gamma}(\theta) \equiv -\log\frac{d\mathcal{P}(\gamma,\theta)}{d\mathcal{P}(\gamma)}(X_n) = -\phi_\gamma(X_n,M_\theta) + \tfrac{1}{2}\|M_\theta\|_\gamma^2,$$

which can be rewritten as

$$\begin{aligned} J_{n,\gamma}(\theta) &= -\phi_\gamma\left(M_\tau + \tfrac{1}{\sqrt{n}}A_{n,\tau}, M_\theta\right) + \tfrac{1}{2}\|M_\theta\|_\gamma^2 \\ &= \tfrac{1}{2}\|M_\theta\|_\gamma^2 - \langle M_\tau, M_\theta\rangle_\gamma - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, M_\theta) \\ &= \tfrac{1}{2}\|M_\theta - M_\tau\|_\gamma^2 - \tfrac{1}{2}\|M_\tau\|_\gamma^2 - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, M_\theta). \end{aligned}$$

It is convenient to define

$$\begin{aligned} L_{n,\gamma}(\theta) &\equiv J_{n,\gamma}(\theta) + \tfrac{1}{2}\|M_\tau\|_\gamma^2 \\ &= \tfrac{1}{2}\|M_\theta - M_\tau\|_\gamma^2 - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, M_\theta). \end{aligned}$$

Of course, the ARE can also be characterized as the minimizer θ of the functional $L_{n,\gamma}(\theta)$. We refer to $L_{n,\gamma}(\theta)$ as the *AR objective functional.*

2.2 Properties of Parametric Estimators

In this section, we present and discuss the main results on the properties of parametric AR estimators. Technical details and proofs are deferred to section 2.6. Under moderate assumptions, Theorem 2.1 establishes the consistency of the first-stage ($i = 1$) estimator in the ARE sequence. This estimator is computed using an arbitrary guess for the covariance parameter. With stronger assumptions, Theorem 2.2 establishes consistency in the regular case. In this context, *regular* means that an estimator is obtained as a zero of the derivative of an objective functional.

Theorem 2.1. *Consider the model of section 2.1. Suppose τ is the true parameter value. Fix $\gamma \in \Theta$. Assume that the following conditions are satisfied.*

(1) Θ *is compact.*

(2) The map $\theta \mapsto M_\theta$ is continuous, so that $L_{n,\gamma}(\theta)$ is continuous on $\theta \in \Theta$ for all n.

(3) For any $\delta > 0$, $\inf\{\|M_\theta - M_\tau\|_\gamma^2 : |\theta - \tau| \geqslant \delta\} > 0$.

(4) $\sup\{|\phi_\gamma(A_{n,\tau}, M_\theta)| : \theta \in \Theta\} = o_p(n^{1/2})$ *as $n \to \infty$, so that*

$$\sup\{|L_{n,\gamma}(\theta) - \mathrm{E}\, L_{n,\gamma}(\theta)| : \theta \in \Theta\} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Then any sequence of AR estimators $\{\tau_n\}_{n\in\mathbb{N}}$ given by

$$L_{n,\gamma}(\tau_n) = \inf\{L_{n,\gamma}(\theta) : \theta \in \Theta\}$$

has the property

$$\tau_n \xrightarrow{p} \tau \quad as \quad n \to \infty.$$

In what follows, the dot denotes differentiation with respect to the parameter. With stronger smoothness conditions, we get:

Theorem 2.2. *Consider the model of section 2.1. Suppose τ is the true parameter value. Fix $\gamma \in \Theta$. Assume that the following conditions are satisfied.*

(1) Θ is compact, and τ is in the interior of Θ.

(2) M_θ is θ-differentiable and $\dot{M}_\theta \in H$, so that $L_{n,\gamma}(\theta)$ is θ-differentiable on Θ for all n.

(3) For any $\delta > 0$, $\inf\{\|M_\theta - M_\tau\|_\gamma^2 : |\theta - \tau| \geqslant \delta\} > 0$.

(4) $\sup\{|\phi_\gamma(A_{n,\tau}, M_\theta)| : \theta \in \Theta\} = o_p(n^{1/2})$ as $n \to \infty$, so that

$$\sup\{|L_{n,\gamma}(\theta) - \mathrm{E}\, L_{n,\gamma}(\theta)| : \theta \in \Theta\} \xrightarrow{p} 0 \quad as \quad n \to \infty.$$

Then, there is a sequence of AR estimators $\{\tau_n\}_{n\in\mathbb{N}}$ satisfying both $\dot{L}_{n,\gamma}(\tau_n) = 0$ and

$$\tau_n \xrightarrow{p} \tau \quad as \quad n \to \infty.$$

We now state two theorems on the asymptotic distributions of AR estimators. We consider the case of a sufficiently differentiable objective with a unique minimum. Of course, these conditions can be weakened in many ways. We restrict our attention to the univariate parameter case, with $d = 1$. The arguments can easily be generalized to the vector case. Theorem 2.3 establishes the asymptotic normality of the first-stage estimator. Theorem 2.4 establishes the asymptotic optimality of AR estimators for $i > 1$. The operator S_{ab}, which appears in the statements of these two theorems, is defined in Lemma 2.8 of section 2.6.1.

Theorem 2.3. *Consider the model of section 2.1. Suppose τ is the true parameter value. Fix $\gamma \in \Theta$. Assume that the following conditions are satisfied.*

(1) Θ is compact, and τ is in the interior of Θ.

(2) M_θ *and* $L_{n,\gamma}(\theta)$ *are twice* θ*-differentiable on* Θ*; and* M_θ, $\dot{M}_\theta$, *and* $\ddot{M}_\theta$ *are in* H_γ *for all* $\theta \in \Theta$.

(3) The maps $\theta \mapsto M_\theta$, $\theta \mapsto \dot{M}_\theta$, *and* $\theta \mapsto \ddot{M}_\theta$, *are continuous; i.e.,* $\|M_\alpha - M_\beta\|_\gamma + \|\dot{M}_\alpha - \dot{M}_\beta\|_\gamma + \|\ddot{M}_\alpha - \ddot{M}_\beta\|_\gamma \to 0$ *if* $|\alpha - \beta| \to 0$.

(4) $\dot{L}_{n,\gamma}^{-1}(0) = \{\tau_n\}$.

(5) $\sup\{|\phi_\gamma(A_\tau, \ddot{M}_\theta - \ddot{M}_\tau)| : |\theta - \tau| \leqslant \varepsilon\} = O_p(1)$ *as* $\varepsilon \to 0$.

(6) For sufficiently small $\varepsilon > 0$, *as* $n \to \infty$,

$$g_\varepsilon(A_{n,\tau}) = \sup_\theta \left\{ |\phi_\gamma(A_{n,\tau}, \ddot{M}_\theta - \ddot{M}_\tau)| : |\theta - \tau| \leqslant \varepsilon \right\} = o_p\left(n^{1/2}\right).$$

Then the AR estimator sequence $\{\tau_n\}_{n\in\mathbb{N}}$ *has the property*

$$\sqrt{n} \cdot (\tau_n - \tau) \xrightarrow{d} Y \sim N\left(0, \frac{\|S_{\gamma\tau}\dot{M}_\tau\|_\gamma^2}{\|\dot{M}_\tau\|_\gamma^4}\right) \quad as \quad n \to \infty.$$

One consequence of the preceding theorem is that

$$\tau_n \xrightarrow{p} \tau \quad \text{as} \quad n \to \infty,$$

so that the first-stage AR estimator is weakly consistent. Also, in the event that $\gamma = \tau$, it is easily shown that

$$\text{Var } \tau_n = \frac{\text{Var } \phi_\tau(\dot{M}_\tau)}{\|\dot{M}_\tau\|_\tau^4} = \frac{1}{\|\dot{M}_\tau\|_\tau^2} = \frac{1}{\text{Var } \phi_\tau(\dot{M}_\tau)}.$$

This is the Cramér-Rao lower bound, assuming its existence. In this case, τ_n is a best asymptotically normal estimator of τ.

With some additional assumptions on the uniform behavior of the parametric function family and the stochastic process, trivial modification of Theorem 2.3 yields the following result on the asymptotic consistency, normality, and optimality of the AR estimators for $i > 1$.

Theorem 2.4. *Consider the model of section 2.1. Let* $N(\tau) \subseteq \Theta$ *be a neighborhood of* τ. *Suppose that the following conditions are satisfied.*

(1) For fixed $X \in \mathcal{S}$ and $F \in H$, the map $\gamma \mapsto \phi_\gamma(X, F)$ is continuous on $N(\tau)$.

(2) For fixed F and $G \in H$, the map $\gamma \mapsto \langle F, G \rangle_\gamma$ is continuous on $N(\tau)$.

(3) The maps $\theta \mapsto M_\theta$, $\theta \mapsto \dot{M}_\theta$, and $\theta \mapsto \ddot{M}_\theta$ are continuous for $\theta \in \Theta$ uniformly for $\gamma \in N(\tau)$; i.e., if $|\alpha - \beta| \to 0$ then

$$\sup \{\|Z_\alpha - Z_\beta\|_\gamma : \gamma \in N(\tau)\} \to 0$$

for each $Z \in \{M, \dot{M}, \ddot{M}\}$.

(4) For sufficiently small $\varepsilon > 0$, as $n \to \infty$,

$$g_\varepsilon(A_{n,\tau}) = \sup_{\gamma \in N(\tau)} \sup_{\theta} \left\{ |\phi_\gamma(A_{n,\tau}, \ddot{M}_\theta - \ddot{M}_\tau)| : |\theta - \tau| \leqslant \varepsilon \right\} = o_p\left(n^{1/2}\right).$$

Then, for any fixed $i > 1$, the AR estimator sequence $\{\tau_{n,i}\}_{n \in \mathbb{N}}$ has the property

$$\sqrt{n}\,(\tau_{n,i} - \tau) \xrightarrow{d} Y_i \sim N\left(0, I(\tau)^{-1}\right) \quad \text{as} \quad n \to \infty,$$

where $I(\tau)$ is Fisher's Information Measure.

This implies that stopping the procedure at any $i \geqslant 2$ gives a best asymptotically normal estimator based on X_n. Recall that the likelihood ratio is

$$\Lambda(Z) = \left[\frac{d\mathcal{P}(\gamma, \theta)}{d\mathcal{P}(\gamma)}(Z)\right]^n = \exp\left[n\phi_\gamma(Z, M_\theta) - \tfrac{n}{2}\|M_\theta\|_\gamma^2\right],$$

and that information is defined using the quantity

$$\frac{d}{d\theta} \log \Lambda(Z) = n\left[\phi_\gamma(Z, \dot{M}_\theta) - \left\langle M_\theta, \dot{M}_\theta \right\rangle_\gamma\right]$$

as follows. Let τ be the true parameter value. Then

$$\mathrm{E}\left[\frac{d}{d\theta} \log \Lambda(X_n)\bigg|_{\theta=\tau}\right] = n\,\mathrm{E}_\tau\,\phi_\gamma(X_n, \dot{M}_\tau) - n\left\langle M_\tau, \dot{M}_\tau \right\rangle_\gamma = 0$$

since $\mathrm{E}_\tau\, \phi_\gamma(Z, F) = \langle \mathrm{E}_\tau\, Z, F \rangle_\gamma$. The information in X_n about τ is defined as

$$I(\tau) = \mathrm{E}\left[\left(\frac{d}{d\theta}\log \Lambda(X_n)\Big|_{\theta=\tau}\right)^2\right] = \mathrm{Var}\left[\frac{d}{d\theta}\log \Lambda(X_n)\Big|_{\theta=\tau}\right].$$

See, for example, Kutoyants [36], p. 10. So here we have

$$\begin{aligned} I(\tau) &= n^2\, \mathrm{Var}\ \phi_\gamma(X_n, \dot{M}_\tau) = n\, \mathrm{Var}\ \phi_\gamma(A_{n,\tau}, \dot{M}_\tau) \\ &\sim n\, \mathrm{Var}\ \phi_\gamma(A_\tau, \dot{M}_\tau) = n\|S_{\gamma\tau}\dot{M}_\tau\|_\gamma^2, \end{aligned}$$

in the sense that $f \sim g$ if and only if $f/g \to 1$ as $n \to \infty$. Thus, the lower bound for the asymptotic variance of an unbiased estimator of τ based on X_n is

$$I(\tau)^{-1} = \frac{1}{n\|S_{\gamma\tau}\dot{M}_\tau\|_\gamma^2},$$

which we may compare to the asymptotic variance

$$\frac{\|S_{\gamma\tau}\dot{M}_\tau\|_\gamma^2}{n\|\dot{M}_\tau\|_\gamma^4}$$

of a first-iteration AR estimator. In the case that $\gamma = \tau$, we simply have

$$I(\tau)^{-1} = \frac{1}{n\|\dot{M}_\tau\|_\tau^2},$$

which is indeed the asymptotic variance of the AR estimators $\tau_{n,i}$ for all $i > 1$, and of $\tau_{n,1}$ when $\tau_{n,0}(= \gamma) = \tau$.

In the remainder of this chapter, we consider some specific applications and develop results concerning the limiting ($i \to \infty$) behavior of the AR estimator sequence for finite samples.

2.3 Application to Density Estimation

Here we develop the form of the AR estimator specific to probability density function (p.d.f.) estimation. The prime denotes differentiation with respect to $t \in$I.

Let $Z_1, \ldots, Z_n$ be i.i.d. real r.v.'s on $I \subseteq \mathbb{R}$ with continuous c.d.f. $F = F_\tau$ where $\tau \in \Theta \subseteq \mathbb{R}^d$ for some feasible parameter set Θ. The empirical cumulative distribution function

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i \leqslant t)$$

has mean-value function

$$\mathrm{E}\, F_n(t) = F(t)$$

and covariance function

$$\mathrm{Cov}[F_n(s), F_n(t)] = \tfrac{1}{n} \left[F(s \wedge t) - F(s)F(t) \right].$$

Suppose $\gamma \in \Theta$. Let $X(t) = X_1(t) = F_n(t) - F_\gamma(t)$, and let $X_0(t)$ be a zero-mean Gaussian process with covariance function $K_\gamma(s,t) = \frac{1}{n}[F_\gamma(s \wedge t) - F_\gamma(s)F_\gamma(t)]$. We model $X(t)$ as a Gaussian process with mean $F_\theta(t) - F_\gamma(t)$ and covariance K_γ.

Let H_γ denote the reproducing kernel Hilbert space with reproducing kernel K_γ. With the map $\phi_\gamma : H_\gamma \to L_2(X)$ given by $\phi_\gamma(K_{\gamma t}) = X(t)$, the required derivative is

$$L(\theta) = \frac{d\mathcal{P}_1}{d\mathcal{P}_0}(X) = \exp\left[\phi_\gamma(F_\theta - F_\gamma) - \tfrac{1}{2} \|F_\theta - F_\gamma\|_\gamma^2 \right],$$

as long as $F_\theta - F_\gamma \in H_\gamma$. We use $F_\theta - F_\gamma$ in the likelihood ratio because H_γ consists of functions A on $[0,1]$ with $A(0) = A(1) = 0$.

Because $K_\gamma(s,t) = u(s \wedge t)\, v(s \vee t)$, where $u = F_\gamma$ and $v = 1 - F_\gamma$, the RKHS inner product for an empirical c.d.f. (Brownian bridge) covariance structure is given by equation (A.2) as

$$\|Z\|_\gamma^2 = \int_0^1 \frac{(Z')^2}{F_\gamma'},$$

as long as $\lim_{x \to 0} Z(x)^2\, F_\gamma(x)^{-1} = 0$ and $\lim_{x \to 1} Z(x)^2 \left[1 - F_\gamma(x)\right]^{-1} = 0$. Then the form of the density functional is

$$J_\gamma(\theta) = -\int \frac{(F_\theta - F_\gamma)'(F_n - F_\gamma)'}{F_\gamma'} + \frac{1}{2} \int_0^1 \frac{[(F_\theta - F_\gamma)']^2}{F_\gamma'}. \qquad (2.3)$$

Some algebra yields

$$J_\gamma(\theta) = -\int \frac{F'_\theta}{F'_\gamma}\, dF_n + \frac{1}{2}\int \frac{(F'_\theta)^2}{F'_\gamma} + \frac{1}{2}.$$

In terms of the densities $F'_\theta = f_\theta$, the AR estimator is that value of θ which minimizes

$$J_\gamma(\theta) = -\int \frac{f_\theta}{f_\gamma}\, dF_n + \frac{1}{2}\int \frac{(f_\theta)^2}{f_\gamma}. \tag{2.4}$$

The asymptotic variance of this AR estimator is

$$\operatorname{Var}\hat{\theta} = \frac{\operatorname{Var}\left[\frac{f_\theta}{f_\gamma}\cdot\frac{d}{d\theta}\log f_\theta\right]}{n\,\mathrm{E}\left[\frac{f_\theta}{f_\gamma}\cdot\left(\frac{d}{d\theta}\log f_\theta\right)^2\right]^2}.$$

Now we characterize the limiting optimization problem associated with this AR estimator sequence in the regular case.

Theorem 2.5. *For arbitrary fixed n, let $\{\theta_i\}_{i\in\mathbb{N}}$ be the sequence of AR estimators for the parameter τ of a random sample density. Suppose the ML and AR estimation problems are regular. If the AR estimator sequence converges, so that*

$$\theta_i \to \theta \quad as \quad i \to \infty,$$

then θ is also the maximum-likelihood estimator of τ.

Theorem 2.5 states that for density estimation based on F_n, the limiting AR estimator is the maximum-likelihood estimator.

2.3.1 Type I Censoring

A data set is said to be *censored* if a known number of observations are missing from one or both ends of the data range. See David [12]. In the case of Type I censoring, we observe the data when they are inside a known range, say $[a, b]$. If a datum falls outside the range, we know which side it is on but not what its value is. So the number of observed data values is itself

a random variable. Formally, we have an underlying sample $Y_1, \ldots, Y_n$ from a distribution with c.d.f. F_θ and p.d.f. $f_\theta = F'_\theta$. However, we do not observe the Y_i directly. Rather, we observe the indicators

$$\delta_i = \begin{cases} -1, & Y_i < a \\ 0, & a \leqslant Y_i \leqslant b \\ 1, & Y_i > b \end{cases}$$

along with the data

$$X_i = \begin{cases} Y_i, & \delta_i = 0 \\ \text{not available}, & \delta_i \neq 0. \end{cases}$$

The corresponding AR objective functional is based on equation (A.2), as is the uncensored objective given by equation (2.4). The result is

$$J_\gamma(\theta) = -\int_a^b \frac{f_\theta}{f_\gamma}\, dF_n - \frac{F_\theta(a)F_n(a)}{F_\gamma(a)} - \frac{F_\theta(b)F_n(b)}{1 - F_\gamma(b)} + \frac{1}{2}\left[\int_a^b \frac{(f_\theta)^2}{f_\gamma} + \frac{F_\theta(a)^2}{F_\gamma(a)} + \frac{F_\theta(b)^2}{1 - F_\gamma(b)}\right].$$

A similar construction also works for general Type I censoring, in which the observable data lie in a union of disjoint intervals. In this case, data are observed in

$$[a_1, b_1] \cup [a_2, b_2] \cup \cdots \cup [a_k, b_k],$$

where $a_1 < b_1 < a_2 < \cdots < b_{k-1} < a_k < b_k$, and data counts are available for each of the complementary regions $(-\infty, a_1)$, (b_1, a_2), $\ldots$, (b_{k-1}, a_k), (b_k, ∞).

Note that censoring is different from *truncation*, in which the distribution itself is modified and the amount of lost data is unknown. Any c.d.f., say F_θ, can be truncated to the interval $[a, b]$. The resulting c.d.f. G_θ is given by

$$G_\theta(x) = \begin{cases} 0, & x < a \\ \frac{F_\theta(x) - F_\theta(a)}{F_\theta(b) - F_\theta(a)}, & a \leqslant x \leqslant b \\ 1, & x > b. \end{cases}$$

2.3.2 Example: Linear Density

Let $X_1, \ldots, X_n$ be i.i.d. on $[0,1]$ with density function $f_b(x) = b + 2(1-b)x$, where $b \in [0,2]$. For fixed f_a, the density functional is

$$J_a(b) = -\int_0^1 \frac{b + 2(1-b)t}{a + 2(1-a)t}\, dF_n(t) + \frac{1}{2}\int_0^1 \frac{[b + 2(1-b)t]^2}{a + 2(1-a)t}\, dt,$$

which is quadratic in b. Therefore, the AR estimator, which is the solution of

$$\underset{b}{\text{minimize}}\ J_a(b) \ \text{ subject to } \ 0 \leqslant b \leqslant 2,$$

can be obtained by "clamping" the minimizer of the unconstrained objective into the feasible region $[0,2]$. So if $\hat{b}$ satisfies $J_a'(\hat{b}) = 0$, then the estimator is

$$b = 2 \wedge (0 \vee \hat{b}) = \min\left(2, \max(0, \hat{b})\right).$$

Differentiating with respect to b, we obtain

$$J_a'(b) = -\int_0^1 \frac{1-2t}{a + 2(1-a)t}\, dF_n(t) + \int_0^1 \frac{(1-2t)\,[b + 2(1-b)t]}{a + 2(1-a)t}\, dt.$$

The solution of $J_a'(b) = 0$ is

$$\hat{b} = \frac{1}{I_2(a)}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1-2X_i}{a + 2(1-a)X_i} - I_1(a)\right],$$

where

$$\begin{aligned} I_1(a) &= \int_0^1 \frac{2t(1-2t)}{a + 2(1-a)t}\, dt = \frac{a[2 - 2a - \log(2-a) + \log a]}{2(1-a)^3} \\ &= \frac{a}{(1-a)^3}\left[1 - a - \frac{1}{2}\log\left(\frac{2}{a} - 1\right)\right] \end{aligned}$$

and

$$\begin{aligned} I_2(a) &= \int_0^1 \frac{(1-2t)^2}{a + 2(1-a)t}\, dt = \frac{2 - 2a - \log(2-a) + \log a}{2(a-1)^3} \\ &= \frac{-1}{(1-a)^3}\left[1 - a - \frac{1}{2}\log\left(\frac{2}{a} - 1\right)\right]. \end{aligned}$$

The unconstrained minimizer is

$$\hat{b} = a + \frac{(a-1)^3}{n\left[1 - a - \frac{1}{2}\log\left(\frac{2}{a} - 1\right)\right]} \sum_{i=1}^{n} \frac{1 - 2X_i}{a + 2(1-a)X_i}, \tag{2.5}$$

which is in fact continuous at $a = 1$ with

$$\lim_{a \to 1} \hat{b} = 1 + \frac{3}{n} \sum_{i=1}^{n} (1 - 2X_i) = 4 - 6\bar{X}.$$

On the other hand, the MLE is the solution $\tilde{b}_n$ of

$$\underset{b}{\text{maximize}} \prod_{i=1}^{n} [b + 2(1-b)X_i] \quad \text{subject to} \quad 0 \leqslant b \leqslant 2, \tag{2.6}$$

which does not exist in closed form. However, Theorem 2.5 implies that the solution of (2.6) may be obtained as the iterated solution of (2.5).

We illustrate these computations with a small Monte-Carlo simulation. For each sample size of $n = 10$, 100, and 1000, we generated 1000 data sets with a "true" parameter value of $b = 0.333$ from the p.d.f. $f(x) = b + 2(1 - b)x$. The ML estimators, $\tilde{b}_n$, were computed using a constrained nonlinear minimization routine in S-PLUS, operating on the negative log-likelihood. The AR estimators, $b_{n,i}$, were computed with a typical initial guess of $b_{n,0} = 1.0$, but occasional initial guesses of 0.5 or 0.1 were required. This happened more often with the smaller data sets ($n = 10$) and hardly at all with the large data sets ($n = 1000$). Mean-squared errors are presented in Table 2.1. We make several comments. For all sample sizes in this simulation, the second-stage ARE has lower error than the corresponding MLE. Theorem 2.5 states that $b_{n,i} \to \tilde{b}_n$ as $i \to \infty$ if the sequence converges. This is true only in the regular cases, where estimators are obtained by zeroing the derivatives of objective functions. In this example, the parameter space is constrained, and sometimes a solution is a boundary point instead of a stationary point. Note that even for small samples, the second-stage ARE is competitive with the MLE.

Table 2.1. Mean-Squared Error for AR and ML Density Estimation Simulation

estimator	n=10	n=100	n=1000
$\tilde{b}_n$	0.1994	0.02006	0.002187
$b_{n,1}$	0.1744	0.01981	0.002590
$b_{n,2}$	0.1851	0.01899	0.002177
$b_{n,3}$	0.1934	0.01997	0.002187
$b_{n,4}$	0.1929	0.01985	0.002187
$b_{n,5}$	0.1951	0.02003	0.002187
$b_{n,6}$	0.1954	0.02000	0.002187
$b_{n,7}$	0.1965	0.02005	0.002187
$b_{n,8}$	0.1966	0.02003	0.002187
$b_{n,9}$	0.1972	0.02006	0.002187
$b_{n,10}$	0.1972	0.02004	0.002187

2.4 Application to Quantile Function Estimation

In this section, we develop the AR estimator for the parameter of a random sample probability law based on the quantile function.

Let $Z_1, \ldots, Z_n$ be i.i.d. real r.v.'s on $I = [0,1]$ with positive density f, (continuous) c.d.f.

$$F(t) = \Pr[Z \leqslant t],$$

quantile function

$$Q(u) = F^{-1}(u) = \inf\{t : F(t) \geqslant u\},$$

and density quantile function

$$g = f \circ Q.$$

Differentiating $F(Q(u)) = u$ yields $Q'(u)\, F'(Q(u)) = 1$, so

$$Q'(u) = \frac{1}{g(u)}.$$

Relevant empirical functions are the empirical distribution $F_n(t)$, the empirical quantile function

$$Q_n(u) = F_n^{-1}(u),$$

and the standardized quantile process

$$V_n(u) = \sqrt{n} g(u)[Q_n(u) - Q(u)] = \sqrt{n} \frac{1}{Q'(u)}[Q_n(u) - Q(u)].$$

Under reasonably mild conditions, the asymptotic distribution of V_n is Gaussian with mean zero and covariance $u \wedge v - uv$. So the asymptotic characteristics of Q_n are

$$\mathrm{E}\, Q_n^*(u) = Q(u) \quad \text{and}$$

$$K(s,t) = \mathrm{Cov}[Q_n^*(s), Q_n^*(t)] = \tfrac{1}{n} Q'(s) Q'(t)(s \wedge t - st),$$

and the asymptotic model is

$$X_n^*(u) = Q_n^*(u) = Q(u) + \tfrac{1}{\sqrt{n}} \cdot Q'(u) B(u),$$

where

$$\mathrm{Cov}[X_n^*(u), X_n^*(v)] = \tfrac{1}{n} Q'(u) Q'(v)(u \wedge v - uv).$$

Since the covariance function can be written as

$$K(s,t) = \tfrac{1}{n} Q'(s \wedge t)(s \wedge t) \cdot Q'(s \vee t)(1 - s \vee t),$$

the corresponding RKHS inner product is given by equation (A.3) as

$$\|Z\|_K^2 = \int_0^1 \left[\left(\frac{Z}{Q'} \right)' \right]^2,$$

as long as $\lim_{x \to 0} x^{-1} Z(x)^2 Q'(x)^{-2} = 0$ and $\lim_{x \to 1} (1-x)^{-1} Z(x)^2 Q'(x)^{-2} = 0$.

In the parametric setting, $Q = Q_\tau$ where $\tau \in \Theta \subseteq \mathbb{R}^d$ for some feasible parameter set Θ. To set up the AR estimation procedure, fix $\gamma \in \Theta$ and let

$X(t) = X_1(t) = Q_n(t) - Q_\gamma(t)$. Let $X_0(t)$ be a zero-mean Gaussian process with covariance function $K_\gamma(s,t) = n^{-1} Q'_\gamma(s) Q'_\gamma(t) \cdot (s \wedge t - st)$. We model $X(t)$ as a Gaussian process with mean $Q_\theta - Q_\gamma$ and covariance K_γ.

AR estimation of θ is accomplished by minimizing the functional

$$J_\gamma(\theta) = -\int_0^1 \left(\frac{Q_n - Q_\gamma}{Q'_\gamma}\right)' \left(\frac{Q_\theta - Q_\gamma}{Q'_\gamma}\right)' + \frac{1}{2}\int_0^1 \left[\left(\frac{Q_\theta - Q_\gamma}{Q'_\gamma}\right)'\right]^2.$$

Multiplying this out and discarding terms constant with respect to θ, we obtain

$$J_\gamma(\theta) = -\int_0^1 \left(\frac{Q_\theta}{Q'_\gamma}\right)' \left(\frac{Q_n}{Q'_\gamma}\right)' + \frac{1}{2}\int_0^1 \left[\left(\frac{Q_\theta}{Q'_\gamma}\right)'\right]^2,$$

which is equivalent for the purposes of optimization. To avoid distributional derivatives, we can integrate by parts to get

$$\int_0^1 \left(\frac{Q_\theta}{Q'_\gamma}\right)' \left(\frac{Q_n}{Q'_\gamma}\right)' = \left(\frac{Q_\theta}{Q'_\gamma}\right)' \frac{Q_n}{Q'_\gamma}\bigg|_0^1 - \int_0^1 \left(\frac{Q_\theta}{Q'_\gamma}\right)'' \frac{Q_n}{Q'_\gamma}.$$

With appropriate behavior at the endpoints, we have

$$J_\gamma(\theta) = \int_0^1 \left(\frac{Q_\theta}{Q'_\gamma}\right)'' \frac{Q_n}{Q'_\gamma} + \frac{1}{2}\int_0^1 \left[\left(\frac{Q_\theta}{Q'_\gamma}\right)'\right]^2.$$

2.4.1 Type II Censoring

Another censoring mechanism is called Type II censoring. Recall (section 2.3.1) in the case of Type I censoring that the observed data are constrained to a certain known range and that the number of observed values is random. The situation is reversed in Type II censoring: a known number of data are excised from each end of the range, and the observed data range is random. Type I censoring is well suited to AR estimation based on the empirical c.d.f. F_n. In a similar fashion, AR estimation based on the empirical quantile function Q_n can be adapted for the Type II censoring model. Parzen [54] points this out for the special case of location and scale estimation.

Of course, if we know the underlying sample size n and the numbers of observations n_a and n_b missing from each end, then we also know the points

p and q which delimit the domain of the corresponding quantile functions. In fact, $p = n_a/n$ and $q = 1 - n_b/n$. For Type II censoring, the AR estimation functional based on Q_n is derived from equation (A.3). The result is

$$J_\gamma(\theta) = -\int_p^q \left(\frac{Q_\theta}{Q'_\gamma}\right)' \left(\frac{Q_n}{Q'_\gamma}\right)' - \frac{1}{p} \cdot \frac{Q_\theta(p)\, Q_n(p)}{Q'_\gamma(p)^2} - \frac{1}{1-q} \cdot \frac{Q_\theta(q)\, Q_n(q)}{Q'_\gamma(q)^2}$$
$$+ \frac{1}{2}\left\{\int_p^q \left[\left(\frac{Q_\theta}{Q'_\gamma}\right)'\right]^2 + \frac{1}{p}\left[\frac{Q_\theta(p)}{Q'_\gamma(p)}\right]^2 + \frac{1}{1-q}\left[\frac{Q_\theta(q)}{Q'_\gamma(q)}\right]^2\right\}.$$

2.4.2 Example: Location and Scale Estimation

Here we consider AR estimators based on the empirical quantile function in the case of location and scale families of distributions. For a fixed density f_o, we have the two-parameter family of distributions

$$f(x;\theta) = \frac{1}{\sigma} f_o\left(\frac{x-\mu}{\sigma}\right),$$

where $\theta = (\mu, \sigma)$. In this case, we show that $\theta_i = \theta_1$ for all $i > 1$.

Other results (see sections 2.3 and 2.5) seem to imply a direct relationship between AR estimators and ML estimators. However, based on the work of Bennett [4], Parzen [54], David [12], and others, it is known that θ_1 need not be the MLE.

To express the location and scale problem in terms of the quantile function, we specify a fixed quantile function Q_o and a two-parameter family of candidate quantile functions

$$\{a + bQ_o : a \in \mathbb{R},\ b > 0\}.$$

For fixed $\gamma = (a_o, b_o)$, we have $Q_\gamma(u) = a_o + b_o Q_o(u)$, and with $\theta = (a, b)$ we have $Q_\theta(u) = a + bQ_o(u)$. The AR estimator θ is the minimizer of

$$J(\theta) = -\int_0^1 \left(\frac{a+bQ_o}{b_o Q'_o}\right)' \left(\frac{Q_n}{b_o Q'_o}\right)' + \frac{1}{2}\int_0^1 \left[\left(\frac{a+bQ_o}{b_o Q'_o}\right)'\right]^2$$

or, equivalently, of

$$J(\theta) = -\int_0^1 \left(\frac{a+bQ_o}{Q_o'}\right)' \left(\frac{Q_n}{Q_o'}\right)' + \frac{1}{2}\int_0^1 \left[\left(\frac{a+bQ_o}{Q_o'}\right)'\right]^2.$$

Thus, the AR estimator is independent of choice of (a_o, b_o).

Let $\langle x, y\rangle = \int_0^1 (x/Q_o')'(y/Q_o')'$, and define the matrices

$$C = \begin{bmatrix} \langle 1, Q_n\rangle \\ \langle Q_o, Q_n\rangle \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} \langle 1,1\rangle & \langle 1, Q_o\rangle \\ \langle 1, Q_o\rangle & \langle Q_o, Q_o\rangle \end{bmatrix}.$$

Then we have

$$\begin{aligned} J(\theta) &= -a\langle 1, Q_n\rangle - b\langle Q_o, Q_n\rangle + \tfrac{1}{2}\left(a^2\langle 1,1\rangle + 2ab\langle 1, Q_o\rangle + b^2\langle Q_o, Q_o\rangle\right) \\ &= -\theta^T C + \tfrac{1}{2}\theta^T R\,\theta, \end{aligned}$$

so the AR estimator θ, which is the minimizer of $J(\theta)$, is given by

$$\theta = R^{-1}\,C.$$

Explicitly, in terms of the components, the estimator is

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \langle 1,1\rangle & \langle 1, Q_o\rangle \\ \langle 1, Q_o\rangle & \langle Q_o, Q_o\rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle 1, Q_n\rangle \\ \langle Q_o, Q_n\rangle \end{bmatrix}.$$

To modify the location and scale problem for Type II censoring (discussed in section 2.4.1), the appropriate inner product is

$$\langle x, y\rangle = \int_p^q \left(\frac{x}{Q_o'}\right)' \left(\frac{y}{Q_o'}\right)' + \frac{1}{p}\cdot\frac{x(p)\,y(p)}{Q_o'(p)^2} + \frac{1}{1-q}\cdot\frac{x(q)\,y(q)}{Q_o'(q)^2}.$$

The resulting estimator is the one obtained by Parzen [54].

2.5 Application to Poisson Process Intensity Estimation

In this section, we develop the AR estimator for the intensity parameter of a completely observed Poisson process with finite mean measure. Again, the prime denotes differentiation with respect to $t \in \mathrm{I}$.

The RKHS inner product for a Poisson process (Wiener process) covariance structure, given by equation (A.1), is

$$\|G\|_K^2 = \int_0^1 \frac{(G')^2}{K'},$$

as long as $\lim_{x \to 0} G(x)^2 \, K(x)^{-1} = 0$. Here, the map ϕ is

$$\phi_K(G) = \int_0^1 \frac{G'}{K'} \, dN,$$

and the AR density functional for a Poisson counting process with mean G and intensity g is then

$$J_\gamma(\theta) = -\int \frac{g_\theta}{g_\gamma} \, dN + \frac{1}{2} \int \frac{(g_\theta)^2}{g_\gamma}.$$

We now characterize the limiting optimization problem associated with this AR estimator sequence in the regular case. The following result is the Poisson process analogue of Theorem 2.5.

Theorem 2.6. *For arbitrary fixed n, let $\{\theta_i\}_{i \in \mathbb{N}}$ be the sequence of AR estimators for the parameter τ of a Poisson process mean measure. Suppose the ML and AR estimation problems are regular. If the AR estimator sequence converges, so that*

$$\theta_i \to \theta \quad as \quad i \to \infty,$$

then θ is the also maximum-likelihood estimator of τ.

In other words, if the AR estimator sequence converges (for any fixed value of n), then it converges to the maximum-likelihood estimator.

2.5.1 Example: Exponential Intensity

Consider a Poisson process on $[0,T]$ with intensity function $\lambda(t) = ae^{-bt}$. The density functional is

$$\begin{aligned} J(a,b) &= -\int_0^T \frac{ae^{-bt}}{a_0 e^{-b_0 t}}\, dN(t) + \frac{1}{2}\int_0^T \frac{\left(ae^{-bt}\right)^2}{a_0 e^{-b_0 t}}\, dt \\ &= -\frac{a}{a_0}\sum_{i=1}^{n} e^{-(b-b_0)t_i} + \frac{a^2}{2a_0}\int_0^T e^{-(2b-b_0)t}\, dt \\ &= -\frac{a}{a_0}\sum_{i=1}^{n} e^{-(b-b_0)t_i} + \frac{a^2}{2a_o(2b-b_0)}\left[1 - e^{-T(2b-b_0)}\right], \end{aligned}$$

as long as $2b - b_0 > 0$. To accomplish the optimization, we differentiate with respect to a,

$$\frac{dJ}{da} = -\frac{1}{a_0}\sum_{i=1}^{n} e^{-(b-b_0)t_i} + \frac{a\left[1 - e^{-T(2b-b_0)}\right]}{a_0(2b-b_0)},$$

and set $dJ/da = 0$ to get

$$a = \frac{2b - b_0}{1 - e^{-T(2b-b_0)}}\sum_{i=1}^{n} e^{-(b-b_0)t_i}. \tag{2.7}$$

Substituting, we have

$$J = \frac{2b-b_0}{a_0\left[1 - e^{-T(2b-b_0)}\right]}\left[\sum_{i=1}^{n} e^{-(b-b_0)t_i}\right]^2 \left[-1 + \frac{1}{2\left[1 - e^{-T(2b-b_0)}\right]}\right].$$

With $T \to \infty$, this becomes

$$J = -\frac{2b-b_0}{2a_0}\left[\sum_{i=1}^{n} e^{-(b-b_0)t_i}\right]^2.$$

So the AR estimator b is the minimizer of

$$J(b) = (b_0 - 2b)\left[\sum_{i=1}^{n} e^{-(b-b_0)t_i}\right]^2,$$

and we recover a from equation (2.7).

$$a = (2b - b_0) \sum_{i=1}^{n} \mathrm{e}^{-(b-b_0)t_i}.$$

On the other hand, the likelihood functional is

$$\begin{aligned} L &= \int_0^T \log \lambda \, dN - \int_0^T \lambda \\ &= \sum_{i=1}^{n} (\log a - b t_i) - a \int_0^T \mathrm{e}^{-bt} \, dt \\ &= n \log a - b \sum_{i=1}^{n} t_i + \frac{a}{b} \left(\mathrm{e}^{-bT} - 1 \right). \end{aligned}$$

So, for large T, we have

$$L = n \log a - b \sum_{i=1}^{n} t_i - \frac{a}{b}.$$

The derivative with respect to a is

$$\frac{dL}{da} = \frac{n}{a} - \frac{1}{b},$$

so that $dL/da = 0$ if and only if $a = nb$. Substituting, we obtain

$$L = n \log nb - b \sum_{i=1}^{n} t_i - n.$$

Now,

$$\frac{dL}{db} = \frac{n}{b} - \sum_{i=1}^{n} t_i,$$

and $dL/db = 0$ if and only if $b^{-1} = n^{-1} \sum_{i=1}^{n} t_i = \bar{t}$. Thus, we finally have the ML estimators

$$\tilde{a} = \frac{n^2}{\sum_{i=1}^{n} t_i} = n \bar{t}^{-1} \quad \text{and} \quad \tilde{b} = \frac{n}{\sum_{i=1}^{n} t_i} = \bar{t}^{-1}.$$

2.6 Technical Details

2.6.1 Distributions of Functionals

The following facts enable us to calculate the asymptotic distributions of functionals that arise in the course of the large-sample analysis of the AR estimator. Here, we make the dependence of ϕ upon the process explicit, as in section 1.4.

Lemma 2.7 is a standard result in Hilbert-space time-series methods. Lemma 2.8 is nonstandard, but required in our applications to characterize the behavior of the map ϕ when the reference covariance structure is different from the true covariance of the process X_n.

Lemma 2.7. *Let X be a stochastic process with sample paths in $\mathcal{S}$, mean-value function M, and covariance function K. Let H_K be the RKHS with reproducing kernel K. Assume that $M \in H_K$. Let $\phi : \mathcal{S} \times H_K \to \mathbb{R}$ be a bilinear functional with $\phi(Y, K_t) = Y(t)$ for all $t \in I$ and $Y \in \mathcal{S}$. Then, for all $f \in H_K$,*

$$\mathrm{E}\,\phi(X, f) = \langle M, f \rangle \quad \textit{and} \quad \mathrm{Var}\,\phi(X, f) = \|f\|^2.$$

Lemma 2.8. *Let X be a stochastic process with sample paths in $\mathcal{S}$, mean-value function M_z, and covariance function K_z under probability law $\mathcal{L}_z$ for $z \in \{a, b\}$. Let H_z be the RKHS with reproducing kernel K_z. Assume that $M_z \in H_z$. Let each bilinear functional $\phi_z : \mathcal{S} \times H_z \to \mathbb{R}$ satisfy $\phi_z(Y, K_{zt}) = Y(t)$ for all $t \in I$ and $Y \in \mathcal{S}$. Suppose that $H_a = H_b = H$, that the linear operator $T = T_{ba}$ on H given by $T(K_{bt}) = K_{at}$ is bounded, and that the linear functional $\mathrm{E}\,\phi_b(X, \cdot)$ on H is bounded. Let $S = S_{ba}$ be a square root of T. Then under law $\mathcal{L}_a$, for all $f \in H$,*

$$\mathrm{E}\,\phi_b(X, f) = \langle M_a, f \rangle_b \quad \textit{and} \quad \mathrm{Var}\,\phi_b(X, f) = \|S_{ba} f\|_b^2.$$

Up to this point, we have not been specific about the definition of convergence in distribution for random functions. There are several different general approaches. See Billingsley [5], Shorack and Wellner [63], and the

references contained in both for discussions of the convergence of probability measures on metric spaces. According to Pollard [56], for example, the typical metric space setup is as follows.

> There are an implicit underlying probability space $(\Omega, \mathcal{F}, \Pr)$, and a function space $\mathcal{S}$ with σ-algebra $\mathcal{A}$ and metric δ. Always, $\mathcal{A} \subseteq \mathcal{B}_\delta(\mathcal{S})$, the Borel σ-algebra generated by the δ-open sets of $\mathcal{S}$. A usual choice is $\mathcal{A} = \mathcal{B}_\delta^o(\mathcal{S})$, the σ-algebra generated by the δ-open balls in $\mathcal{S}$. The random elements $Z : \Omega \to \mathcal{S}$ are then the $\mathcal{F}/\mathcal{A}$-measurable functions. Equip the reals with the usual Borel σ-algebra $\mathcal{B}$. Let $f : \mathcal{S} \to \mathbb{R}$ be bounded, δ-continuous, and $\mathcal{A}/\mathcal{B}$-measurable. Then, $Z_n \xrightarrow{d} Z$ in $\mathcal{S}$ means that $f(Z_n) \xrightarrow{d} f(Z)$ for each such f.

For a different (metric-free) approach, see Kallenberg [30], Karr [31], and their references. In any case, however, the useful and practical application of convergence in distribution can always be described as:

> Given a sequence Z_n of random elements taking values in a space $\mathcal{S}$ and a functional $f : \mathcal{S} \to \mathbb{R}$, one seeks to establish conditions that imply that the image of f converges in (one-dimensional) distribution in $\mathbb{R}$; i.e.,
>
> $$f(Z_n) \xrightarrow{d} f(Z) \text{ as } n \to \infty. \tag{2.8}$$
>
> In general, this can be accomplished by identifying an appropriate mode of convergence in distribution in $\mathcal{S}$; i.e.,
>
> $$Z_n \xrightarrow{d} Z \text{ as } n \to \infty, \tag{2.9}$$
>
> and then showing that f belongs to a class of functions for which (2.9) implies (2.8).

We are not concerned with the *theoretical* details but need only consider the *practical* use of the technology, since our interest is in the application of the AR concept to specific stochastic processes X_n.

Therefore, throughout this work we assume that the convergence in distribution of the stochastic processes specified in statements (1.2) and (2.1); i.e.,

$$A_{n,\tau} \xrightarrow{d} A_\tau \quad \text{as} \quad n \to \infty, \tag{2.10}$$

implies convergence in distribution of the functionals

$$\phi_\tau(A_{n,\tau}, G) \xrightarrow{d} \phi_\tau(A_\tau, G) \quad \text{as} \quad n \to \infty \tag{2.11}$$

for any $G \in H_\tau$.

While the manner of convergence in (2.10) for each of the processes we consider may be different (e.g., in the metric space setting, the σ-algebra $\mathcal{A}$ and metric δ are specific to the application), statement (2.11) holds nonetheless.

Now we consider the distribution of the limit $\phi_\tau(A_\tau, G)$ in (2.11). It is easy to show that this quantity has a normal distribution. This follows from the fact that the set $\left\{\sum_{i=1}^n a_i K_{t_i} : n \in \mathbb{N},\, t_i \in I,\, a_i \in \mathbb{R}\right\}$ is dense in H, upon consideration of a Cauchy sequence.

The large-sample properties of AR estimators developed in this chapter are determined by the asymptotic behavior of $\phi_\gamma(A_{n,\tau}, G)$ for various values of G. In particular, suppose that M_θ, $\dot{M}_\theta$, and $\ddot{M}_\theta$ lie in H_γ for all $\theta \in \Theta$, where the dot denotes differentiation with respect to θ. In light of the previous development, the asymptotic distributions of functionals of interest are

$$\begin{aligned}
&\phi_\gamma(A_{n,\tau}, M_\theta) \xrightarrow{d} \phi_\gamma(A_\tau, M_\theta) \sim N(0, \|S_{\gamma\tau} M_\theta\|_\gamma^2), \\
&\phi_\gamma(A_{n,\tau}, \dot{M}_\theta) \xrightarrow{d} \phi_\gamma(A_\tau, \dot{M}_\theta) \sim N(0, \|S_{\gamma\tau} \dot{M}_\theta\|_\gamma^2), \quad \text{and} \\
&\phi_\gamma(A_{n,\tau}, \ddot{M}_\theta) \xrightarrow{d} \phi_\gamma(A_\tau, \ddot{M}_\theta) \sim N(0, \|S_{\gamma\tau} \ddot{M}_\theta\|_\gamma^2).
\end{aligned}$$

The AR objective functional and its θ-derivatives are

$$\begin{aligned}
L_{n,\gamma}(\theta) &= \tfrac{1}{2}\|M_\theta - M_\tau\|_\gamma^2 - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, M_\theta), \\
\dot{L}_{n,\gamma}(\theta) &= \left\langle M_\theta - M_\tau, \dot{M}_\theta \right\rangle_\gamma - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, \dot{M}_\theta), \quad \text{and} \\
\ddot{L}_{n,\gamma}(\theta) &= \left\langle M_\theta - M_\tau, \ddot{M}_\theta \right\rangle_\gamma + \|\dot{M}_\theta\|_\gamma^2 - \tfrac{1}{\sqrt{n}}\phi_\gamma(A_{n,\tau}, \ddot{M}_\theta).
\end{aligned}$$

The asymptotic distributions of these quantities follow from

$$\begin{aligned}
\sqrt{n}\cdot\left[L_{n,\gamma}(\theta)-\tfrac{1}{2}\|M_\theta-M_\tau\|_\gamma^2\right] &= -\phi_\gamma(A_{n,\tau},M_\theta),\\
\sqrt{n}\cdot\left[\dot{L}_{n,\gamma}(\theta)-\left\langle M_\theta-M_\tau,\dot{M}_\theta\right\rangle_\gamma\right] &= -\phi_\gamma(A_{n,\tau},\dot{M}_\theta), \quad \text{and}\\
\sqrt{n}\cdot\left[\ddot{L}_{n,\gamma}(\theta)-\left\langle M_\theta-M_\tau,\ddot{M}_\theta\right\rangle_\gamma-\|\dot{M}_\theta\|_\gamma^2\right] &= -\phi_\gamma(A_{n,\tau},\ddot{M}_\theta).
\end{aligned}$$

In particular, with $\theta=\tau$, these become

$$\begin{aligned}
\sqrt{n}L_{n,\gamma}(\tau) &= -\phi_\gamma(A_{n,\tau},M_\tau),\\
\sqrt{n}\dot{L}_{n,\gamma}(\tau) &= -\phi_\gamma(A_{n,\tau},\dot{M}_\tau), \quad \text{and}\\
\sqrt{n}\cdot\left[\ddot{L}_{n,\gamma}(\tau)-\|\dot{M}_\tau\|_\gamma^2\right] &= -\phi_\gamma(A_{n,\tau},\ddot{M}_\tau).
\end{aligned}$$

More details about the behavior of functionals of the observed process and the AR estimators are developed as needed in the proofs of the lemmas, theorems, and corollaries of chapter 2. These proofs are presented collectively in the next section.

2.6.2 Proofs

Proof of Theorem 2.1. Since Var $L_{n,\gamma}(\theta)\to 0$ as $n\to\infty$, Chebychev's inequality gives convergence in probability of $L_{n,\gamma}(\theta)$ to its expectation $\frac{1}{2}\|M_\theta-M_\tau\|_\gamma^2$. In particular, $L_{n,\gamma}(\tau)\xrightarrow{p}0$ as $n\to\infty$. Now fix $\delta>0$ and let

$$\varepsilon=\inf\{\mathrm{E}\,L_{n,\gamma}(\theta):|\theta-\tau|\geqslant\delta\}\,.$$

Pick any $\alpha<\varepsilon/2$. Define the events $A_0=\{|\tau_n-\tau|<\delta\}$ and $A_1=\{L_{n,\gamma}(\tau_n)<\alpha\}$. For large enough n, we have

$$\begin{aligned}
\Pr(A_1A_0^c) &= \Pr(L_{n,\gamma}(\tau_n)<\alpha \text{ and } |\tau_n-\tau|\geqslant\delta)\\
&\leqslant \Pr(\sup\{|L_{n,\gamma}(\theta)-\mathrm{E}\,L_{n,\gamma}(\theta)|:|\theta-\tau|\geqslant\delta\}>\alpha)\\
&\leqslant \Pr(\sup\{|L_{n,\gamma}(\theta)-\mathrm{E}\,L_{n,\gamma}(\theta)|:\theta\in\Theta\}>\alpha)\\
&< \alpha
\end{aligned}$$

and

$$\Pr(A_1) = \Pr(L_{n,\gamma}(\tau_n) < \alpha) \geqslant \Pr(L_{n,\gamma}(\tau) < \alpha) \geqslant 1 - \alpha.$$

Partitioning on A_0, we have

$$1 - \alpha \leqslant \Pr(A_1) = \Pr(A_1 A_0) + \Pr(A_1 A_0^c) < \Pr(A_0) + \alpha.$$

Putting this together, we have $\Pr(A_0) > 1 - 2\alpha$, as required. □

Proof of Theorem 2.2. For small enough δ, the ball $\{\theta : |\theta - \tau| < \delta\}$ lies in the interior of Θ. So any minimizer τ_n of $L_{n,\gamma}$ with $|\tau_n - \tau| < \delta$ also has $\dot{L}(\tau_n) = 0$. The conclusion follows by Theorem 2.1. □

Proof of Theorem 2.3. Sen and Singer [61] use the following approach to prove a theorem about the properties of maximum-likelihood estimators. They attribute the technique to Le Cam [42], Hájek [25], and Inagaki [28].

Fix an arbitrary $K \in (0, \infty)$, and for $|u| \leqslant K$ consider the Taylor's series expansion of $L_{n,\gamma}$ about τ.

$$L_{n,\gamma}(\tau + \tfrac{1}{\sqrt{n}}u) = L_{n,\gamma}(\tau) + \tfrac{1}{n}\lambda_n(u),$$

where

$$\lambda_n(u) = \sqrt{n}u\dot{L}_{n,\gamma}(\tau) + \tfrac{1}{2}u^2\ddot{L}_{n,\gamma}(\tau_{n*}) \quad \text{and} \quad \tau_{n*}(u) \in (\tau, \tau + n^{-1/2}u),$$

so that $|\tau_{n*} - \tau| < n^{-1/2}|u| \leqslant n^{-1/2}K$. Let

$$Z_n(u) = \ddot{L}_{n,\gamma}(\tau_{n*}) - \ddot{L}_{n,\gamma}(\tau) = B_n(u) + C_n(u) - n^{-1/2}D_n(u),$$

where

$$\begin{aligned} B_n(u) &= \left\langle M_{\tau_{n*}} - M_\tau, \ddot{M}_{\tau_{n*}} \right\rangle_\gamma, \\ C_n(u) &= \|\dot{M}_{\tau_{n*}}\|_\gamma^2 - \|\dot{M}_\tau\|_\gamma^2, \text{ and} \\ D_n(u) &= \phi_\gamma(A_{n,\tau}, \ddot{M}_{\tau_{n*}} - \ddot{M}_\tau). \end{aligned}$$

Then we have

$$\lambda_n(u) = \sqrt{n}u\dot{L}_{n,\gamma}(\tau) + \tfrac{1}{2}u^2\ddot{L}_{n,\gamma}(\tau) + \tfrac{1}{2}u^2\left[B_n(u) + C_n(u) + \tfrac{1}{\sqrt{n}}D_n(u)\right].$$

For any $\varepsilon > 0$, fix an $n_0 > K^2/\varepsilon^2$ and note that if $n > n_0$, then $n^{-1/2}K < \varepsilon$ and $|\tau_{n*} - \tau| < \varepsilon$. Therefore, for all sufficiently large n, we have

$$\begin{aligned}\sup_u\{|B_n(u)| : |u| \leqslant K\} \leqslant B_\varepsilon &= \sup_\theta\left\{\left|\left\langle M_\theta - M_\tau, \ddot{M}_\theta\right\rangle_\gamma\right| : |\theta - \tau| \leqslant \varepsilon\right\} \\ &\leqslant \sup_\theta\left\{\|M_\theta - M_\tau\|_\gamma \cdot \|\ddot{M}_\theta\|_\gamma : |\theta - \tau| \leqslant \varepsilon\right\}.\end{aligned}$$

Since the continuity assumptions imply that $B_\varepsilon \to 0$ as $\varepsilon \to 0$, we have $|B_n(u)| \xrightarrow{a.s.} 0$ as $n \to \infty$ uniformly in $|u| \leqslant K$. Likewise, for all large enough n,

$$\sup_u\{|C_n(u)| : |u| \leqslant K\} \leqslant C_\varepsilon = \sup_\theta\left\{\left|\|\dot{M}_\theta\|_\gamma^2 - \|\dot{M}_\tau\|_\gamma^2\right| : |\theta - \tau| < \varepsilon\right\},$$

and $C_\varepsilon \to 0$ as $\varepsilon \to 0$. Therefore, $|C_n(u)| \xrightarrow{a.s.} 0$ as $n \to \infty$ uniformly in $|u| \leqslant K$. Again, for all sufficiently large n,

$$\sup_u\{|D_n(u)| : |u| \leqslant K\} \leqslant \sup_\theta\{|\phi_\gamma(A_{n,\tau}, \ddot{M}_\theta - \ddot{M}_\tau)| : |\theta - \tau| \leqslant \varepsilon\} = g_\varepsilon(A_{n,\tau}).$$

Since $g_\varepsilon(A_{n,\tau}) = o_p(n^{1/2})$, we have $\sup_u\{|D_n(u)| : |u| \leqslant K\} = o_p(n^{1/2})$, and therefore $\sup_u\{n^{-1/2}|D_n(u)| : |u| \leqslant K\} = o_p(1)$ as $n \to \infty$. Since we also have $\ddot{L}_{n,\gamma}(\tau) \xrightarrow{p} \|\dot{M}_\tau\|_\gamma^2$, we may conclude that uniformly on $|u| \leqslant K$

$$\lambda_n(u) = \sqrt{n}u\dot{L}_{n,\gamma}(\tau) + \tfrac{1}{2}u^2\|\dot{M}_\tau\|_\gamma^2 + o_p(1) \quad \text{as} \quad n \to \infty.$$

With $u_0 = -\sqrt{n}\dot{L}_{n,\gamma}(\tau)/\|\dot{M}_\tau\|_\gamma^2$ and $c = \|\dot{M}_\tau\|_\gamma^2/2$, we can write

$$\lambda_n(u) = c(u - u_0)^2 - cu_0^2 + g_n(u),$$

where $\sup\{|g_n(u)| : |u| \leqslant K\} \to 0$ as $n \to \infty$.

Define the events $A \equiv |u_0| < K - \alpha$ and $B \equiv |u_n - u_0| < \alpha$. For any small $\alpha > 0$, choose K and N large enough so that $\Pr(A) > 1 - \alpha$

and $\Pr\left(\sup\{|g_n(u)| : |u| < K\} < \frac{1}{2}c\alpha^2\right) > 1-\alpha$ if $n > N$. Let $\lambda_n(u_n) = \inf\{\lambda_n(u) : |u| \leqslant K\}$. Conditional on A, we have

$$\begin{aligned} B &\Longleftarrow \lambda_n(u_0) < \inf\{\lambda_n(u) : |u-u_0| \geqslant \alpha\} \\ &\Longleftarrow -cu_0^2 + \sup_{|u|<K} |g_n(u)| < c\alpha^2 - cu_0^2 - \sup_{|u|<K} |g_n(u)| \\ &\Longleftrightarrow \sup_{|u|<K} |g_n(u)| < \tfrac{1}{2}c\alpha^2; \end{aligned}$$

and since $\Pr(B) \geqslant \Pr(A \cap B) = \Pr(A)\Pr(B|A) > (1-\alpha)^2$, we have $\Pr(B) \to 1$ as $\alpha \to 0$. Therefore $u_n - u_0 \xrightarrow{p} 0$ as $n \to \infty$. But $\tau_n = \tau + n^{-1/2}u_n$, so

$$\sqrt{n}(\tau_n - \tau) + \frac{\sqrt{n}\dot{L}_{n,\gamma}(\tau)}{\|\dot{M}_\tau\|_\gamma^2} \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.$$

Since

$$\sqrt{n}\dot{L}_{n,\gamma}(\tau) = -\phi_\gamma(A_{n,\tau}, \dot{M}_\tau) \xrightarrow{d} -\phi_\gamma(A_\tau, \dot{M}_\tau) \sim N(0, \|S_{\gamma\tau}\dot{M}_\tau\|_\gamma^2)$$

as $n \to \infty$, the result is established. □

Proof of Theorem 2.4. By Theorem 2.3, we have

$$\sqrt{n} \cdot (\tau_{n,1} - \tau) \xrightarrow{d} Y_1 \sim N\left(0, \frac{\|S_{\tau_{n,0}\tau}\dot{M}_\tau\|_{\tau_{n,0}}^2}{\|\dot{M}_\tau\|_{\tau_{n,0}}^4}\right) \quad \text{as} \quad n \to \infty,$$

so that, in particular, $\tau_{n,1} \xrightarrow{p} \tau$ as $n \to \infty$. Thus, for arbitrarily small $\beta > 0$, there is an N such that if $n > N$ then $\Pr(\tau_{n,1} \in N(\tau)) > 1-\beta$. Now let $i = 2$. Fix an arbitrary $K \in (0, \infty)$. Proceeding as in the proof of Theorem 2.3, we can write the Taylor's series expansion of $L_{n,\tau_{n,1}}$ about τ for $|u| < K$.

$$L_{n,\tau_{n,1}}(\tau + \tfrac{1}{\sqrt{n}}u) = L_{n,\tau_{n,1}}(\tau) + \tfrac{1}{n}\lambda_n(u),$$

where

$$\begin{aligned} \lambda_n(u) &= \sqrt{n}u\dot{L}_{n,\tau_{n,1}}(\tau) + \tfrac{1}{2}u^2\ddot{L}_{n,\tau_{n,1}}(\tau_{n*}) \\ &= \sqrt{n}u\dot{L}_{n,\tau_{n,1}}(\tau) + \tfrac{1}{2}u^2\ddot{L}_{n,\tau_{n,1}}(\tau) + \tfrac{1}{2}u^2\left[B_n(u) + C_n(u) + \tfrac{1}{\sqrt{n}}D_n(u)\right], \end{aligned}$$

in which

$$B_n(u) = \left\langle M_{\tau_{n*}} - M_\tau, \ddot{M}_{\tau_{n*}} \right\rangle_{\tau_{n,1}},$$
$$C_n(u) = \|\dot{M}_{\tau_{n*}}\|^2_{\tau_{n,1}} - \|\dot{M}_\tau\|^2_{\tau_{n,1}}, \text{ and}$$
$$D_n(u) = \phi_{\tau_{n,1}}(A_{n,\tau}, \ddot{M}_{\tau_{n*}} - \ddot{M}_\tau),$$

with $\tau_{n*}(u) = \tau_{n*} \in (\tau, \tau + n^{-1/2}u)$, so that $|\tau_{n*} - \tau| < n^{-1/2}|u| \leqslant n^{-1/2}K$. For sufficiently large n, we have

$$\sup_u\{|B_n(u)| : |u| \leqslant K\} \leqslant \sup_\theta \left\{ \left| \left\langle M_\theta - M_\tau, \ddot{M}_\theta \right\rangle_{\tau_{n,1}} \right| : |\theta - \tau| \leqslant \varepsilon \right\}$$
$$\leqslant \sup_\theta \left\{ \|M_\theta - M_\tau\|_{\tau_{n,1}} \cdot \|\ddot{M}_\theta\|_{\tau_{n,1}} : |\theta - \tau| \leqslant \varepsilon \right\},$$

and therefore with

$$B_\varepsilon = \sup_{\gamma \in N(\tau)} \sup_\theta \left\{ \|M_\theta - M_\tau\|_\gamma \cdot \|\ddot{M}_\theta\|_\gamma : |\theta - \tau| \leqslant \varepsilon \right\}$$

we have

$$\Pr\left[\sup_u\{|B_n(u)| : |u| \leqslant K\} \leqslant B_\varepsilon \right] > 1 - \beta.$$

The continuity assumptions imply that $B_\varepsilon \to 0$ as $\varepsilon \to 0$, so we have $|B_n(u)| \xrightarrow{p} 0$ as $n \to \infty$ uniformly in $|u| \leqslant K$. Likewise, for all large enough n,

$$\sup_u\{|C_n(u)| : |u| \leqslant K\} \leqslant \sup_\theta \left\{ \left| \|\dot{M}_\theta\|^2_{\tau_{n,1}} - \|\dot{M}_\tau\|^2_{\tau_{n,1}} \right| : |\theta - \tau| < \varepsilon \right\}.$$

So with

$$C_\varepsilon = \sup_{\gamma \in N(\tau)} \sup_\theta \left\{ \left| \|\dot{M}_\theta\|^2_\gamma - \|\dot{M}_\tau\|^2_\gamma \right| : |\theta - \tau| < \varepsilon \right\},$$

we have

$$\Pr\left[\sup_u\{|C_n(u)| : |u| \leqslant K\} \leqslant C_\varepsilon \right] > 1 - \beta.$$

The continuity assumptions imply that $C_\varepsilon \to 0$ as $\varepsilon \to 0$, so we have $|C_n(u)| \xrightarrow{p} 0$ as $n \to \infty$ uniformly in $|u| \leqslant K$. Again, for all sufficiently large n,

$$\sup_u\{|D_n(u)| : |u| \leqslant K\} \leqslant \sup_\theta\{|\phi_{\tau_{n,1}}(A_{n,\tau}, \ddot{M}_\theta - \ddot{M}_\tau)| : |\theta - \tau| \leqslant \varepsilon\},$$

so

$$\Pr\left[\sup_u\{|D_n(u)| : |u| \leqslant K\} < g_\varepsilon(A_{n,\tau})\right] > 1-\beta.$$

Since $g_\varepsilon(A_{n,\tau}) = o_p(n^{1/2})$, we have $\sup_u\{|D_n(u)| : |u| \leqslant K\} = o_p(n^{1/2})$, and therefore $\sup_u\{n^{-1/2}|D_n(u)| : |u| \leqslant K\} = o_p(1)$ as $n \to \infty$. The γ-continuity of ϕ_γ implies that

$$\phi_{\tau_{n,1}}(A_{n,\tau}, F) - \phi_\tau(A_{n,\tau}, F) \xrightarrow{p} 0,$$

and it is basic that

$$\tfrac{1}{\sqrt{n}}\phi_\tau(A_{n,\tau}, F) \xrightarrow{p} 0$$

for $F = M_\tau$, $\dot{M}_\tau$, or $\ddot{M}_\tau$. Therefore,

$$\ddot{L}_{n,\tau_{n,1}}(\tau) - \|\dot{M}_\tau\|^2_{\tau_{n,1}} = -\tfrac{1}{\sqrt{n}}\phi_{\tau_{n,1}}(A_{n,\tau}, \ddot{M}_\tau) \xrightarrow{p} 0,$$

and we may conclude that uniformly on $|u| \leqslant K$,

$$\lambda_n(u) = \sqrt{n}u\dot{L}_{n,\tau_{n,1}}(\tau) + \tfrac{1}{2}u^2\|\dot{M}_\tau\|^2_{\tau_{n,1}} + o_p(1) \text{ as } n \to \infty.$$

Since $\dot{L}_{n,\gamma}(\tau) = -n^{-1/2}\phi_\gamma(A_{n,\tau}, \dot{M}_\tau)$, and $\langle\cdot,\cdot\rangle_\gamma$ is γ-continuous, we in fact have

$$\lambda_n(u) = \sqrt{n}u\dot{L}_{n,\tau}(\tau) + \tfrac{1}{2}u^2\|\dot{M}_\tau\|^2_\tau + o_p(1) \text{ as } n \to \infty$$

uniformly for $|u| < K$. As in the proof of the previous theorem, we obtain

$$\sqrt{n}(\tau_{n,2} - \tau) + \frac{\sqrt{n}\dot{L}_{n,\tau}(\tau)}{\|\dot{M}_\tau\|^2_\tau} \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Since

$$\sqrt{n}\dot{L}_{n,\tau}(\tau) = -\phi_\tau(A_{n,\tau}, \dot{M}_\tau) \xrightarrow{d} -\phi_\tau(A_\tau, \dot{M}_\tau) \sim N(0, \|\dot{M}_\tau\|^2_\tau),$$

we have established that

$$\sqrt{n}(\tau_{n,2} - \tau) \xrightarrow{d} Y_2 \sim N\left(0, \frac{1}{\|\dot{M}_\tau\|^2_\tau}\right) \text{ as } n \to \infty.$$

By induction, we may conclude that

$$\sqrt{n}(\tau_{n,i} - \tau) \xrightarrow{d} Y_i \sim N\left(0, \frac{1}{\|\dot{M}_\tau\|^2_\tau}\right) \text{ as } n \to \infty$$

for all $i > 1$. □

Proof of Theorem 2.5. If the AR estimator sequence converges to some value θ, then θ is a fixed point of the operator S_n defined on page 20. So $S_n(\theta) = \theta$, which means that

$$\frac{d}{d\xi} J_\theta(\xi)\Big|_{\xi=\theta} = 0.$$

Differentiating, we have

$$\frac{d}{d\xi} J_\theta(\xi) = -\int \frac{\frac{d}{d\xi} f_\xi}{g_\theta}\, dF_n + \int \frac{f_\xi \frac{d}{d\xi} f_\xi}{f_\theta}.$$

Substituting $\xi = \theta$, we get

$$\begin{aligned}
0 &= -\int \frac{\frac{d}{d\xi} f_\xi|_{\xi=\theta}}{f_\theta}\, dF_n + \int \frac{d}{d\xi} f_\xi|_{\xi=\theta} \\
&= -\int \frac{d}{d\xi} \log f_\xi|_{\xi=\theta}\, dF_n + \int \frac{d}{d\xi} f_\xi|_{\xi=\theta} \\
&= \frac{d}{d\xi}\left[-\int \log f_\xi\, dF_n + \int f_\xi\right]\Big|_{\xi=\theta} \\
&= \frac{d}{d\xi}\left[-\int \log f_\xi\, dF_n + 1\right]\Big|_{\xi=\theta} \\
&= -\frac{d}{d\xi} \int \log f_\xi\, dF_n.
\end{aligned}$$

Therefore, θ solves

$$\underset{\xi}{\text{maximize}} \int \log f_\xi\, dF_n,$$

which characterizes the maximum-likelihood estimator of a random sample density parameter. □

Proof of Theorem 2.6. If the AR estimator sequence converges to some value θ, then θ is a fixed point of the operator S_n defined on page 20. So $S_n(\theta) = \theta$, which means that

$$\frac{d}{d\xi} J_\theta(\xi)\Big|_{\xi=\theta} = 0.$$

Differentiating, we have

$$\frac{d}{d\xi} J_\theta(\xi) = -\int \frac{\frac{d}{d\xi} g_\xi}{g_\theta}\, dN + \int \frac{g_\xi \frac{d}{d\xi} g_\xi}{g_\theta}.$$

Substituting $\xi = \theta$, we get

$$\begin{aligned}
0 &= -\int \frac{\frac{d}{d\xi} g_\xi|_{\xi=\theta}}{g_\theta}\, dN + \int \frac{d}{d\xi} g_\xi|_{\xi=\theta} \\
&= -\int \frac{d}{d\xi} \log g_\xi|_{\xi=\theta}\, dN + \int \frac{d}{d\xi} g_\xi|_{\xi=\theta} \\
&= \frac{d}{d\xi} \left[-\int \log g_\xi\, dN + \int_0^1 g_\xi \right] \Bigg|_{\xi=\theta} \\
&= \frac{d}{d\xi} \left[-\int \log g_\xi\, dN + G_\xi(1) \right] \Bigg|_{\xi=\theta}.
\end{aligned}$$

Therefore, θ solves

$$\underset{\xi}{\text{maximize}} \int \log g_\xi\, dN - G_\xi(1),$$

which characterizes the maximum-likelihood estimator of a Poisson process intensity parameter. □

Proof of Lemma 2.7. See Parzen [51]. □

Proof of Lemma 2.8. We suppress the dependence of ϕ on $\mathcal{S}$ and write $\phi(f)$ for $\phi(X, f)$.

Since $\mathrm{E}\,\phi_b$ is a continuous linear functional on H, there is a unique $s \in H$ with $\mathrm{E}\,\phi_b(f) = \langle s, f \rangle_b$ for all $f \in H$. In particular, $\mathrm{E}\,\phi_b(K_{bt}) = \langle s, K_{bt} \rangle_b = s(t)$. On the other hand, we have $\mathrm{E}\,\phi_b(K_{bt}) = \mathrm{E}\,X(t) = M_a(t)$. Therefore, $s(t) = M_a(t)$, and $\mathrm{E}\,\phi_b(f) = \langle M_a, f \rangle_b$ as required.

Consider the bounded linear operator $T = T_{ba}$ on H given by $(Tf)(x) = \langle f, K_{ax} \rangle_b$, and the associated bounded bilinear functional $B(f, g) = \langle Tf, g \rangle_b$. Note that

$$(TK_{bt})(x) = \langle K_{bt}, K_{ax} \rangle_b = K_{ax}(t) = K_{at}(x).$$

Also,

$$B(K_{bs}, K_{bt}) = \langle TK_{bs}, K_{bt} \rangle_b = \langle K_{as}, K_{bt} \rangle_b = K_{as}(t) = K_a(s,t).$$

But since

$$K_a(s,t) = \mathrm{Cov}[X(s), X(t)] = \mathrm{Cov}[\phi_b(K_{bs}), \phi_b(K_{bt})],$$

we have in general $B(f,g) = \mathrm{Cov}[\phi_b(f), \phi_b(g)]$ and $B(f,f) = \mathrm{Var}\ \phi_b(f)$. Since B is positive definite, symmetric, and self-adjoint, we know that T has a positive definite, symmetric, self-adjoint square root $S = S_{ba}$. So we can write $\langle Tf, f \rangle_b = \langle Sf, Sf \rangle_b = \|Sf\|_b^2$, whence in fact

$$\mathrm{Var}\ \phi_b(f) = \|S_{ba} f\|_b^2 = \left\langle \left\langle f, K_{a(\cdot)} \right\rangle_b, f(\cdot) \right\rangle_b$$

for all $f \in H$. □

3. Nonparametric AR Estimation

In this chapter, we consider asymptotic regression estimation in the setting of an infinite-dimensional real parameter space. As one would expect in this case, even for a fixed initial guess γ, the minimization problem either has no solution or results in an estimator that is not smooth enough. In other words, the problem is ill-posed. However, we show that the technique of regularization, or penalization, can be applied to these problems to produce estimators that have desirable asymptotic properties.

We address nonparametric estimation in the context of specific applications. Basic density estimation is the subject of section 3.1. Density estimation for inverse problems is considered in section 3.2. Section 3.3 contains a short discussion of the application of ARE to Poisson process intensity estimation. Proofs are deferred to section 3.4.

3.1 Density Estimation

Let $X_1, \ldots, X_n$ be i.i.d. random variables on $I \subseteq \mathbb{R}$ with c.d.f. F_o and p.d.f. $F_o' = f_o$. Let h be a p.d.f. on I. In what follows, $\int$ means $\int_I$. With the objective functional

$$J_{n,h}^*(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2}\int \frac{f^2}{h}$$

and constraint set $\mathcal{C} = \{f : f \geqslant 0,\ \int f = 1\}$, the natural nonparametric version of the AR density estimation problem is

$$\underset{f}{\text{minimize}}\ J_{n,h}^*(f) \ \text{ subject to } \ f \in L_2 \cap \mathcal{C}. \tag{3.1}$$

However, this problem is ill-posed in the following sense.

Theorem 3.1. *Problem* (3.1) *has no solution.*

In the proof, we construct a sequence $\{g_m\}_{m=1}^{\infty}$ in $L_2 \cap \mathcal{C}$ with the property that $J^*(g_m) \to -\infty$ as $m \to \infty$, thereby showing that J^* is unbounded.

In fact, $G_m(t) = \int_{x \leqslant t} g_m(x)\,dx$ converges in L_2 to F_n, the empirical c.d.f. Since the optimization problem (3.1) has no solution in L_2, it is useless for estimating f_o and its derivatives.

The same thing happens when one attempts to extend maximum likelihood density estimation by relaxing the parametric restriction. The problem

$$\underset{f}{\text{maximize}} \prod_{i=1}^{n} f(X_i) \quad \text{subject to} \quad f \in L_2 \cap \mathcal{C} \tag{3.2}$$

has no solution. The proof of this fact utilizes the same construction as Theorem 3.1. See Thompson and Tapia [73].

One approach that has been exploited is to replace problem (3.2) by an approximate version that has a useful solution—namely,

$$\underset{f}{\text{maximize}} \prod_{i=1}^{n} f(X_i) \exp[-n\alpha\nu(f)] \quad \text{subject to} \quad f \in L_2 \cap \mathcal{C}. \tag{3.3}$$

Note that equivalent formulations of problems (3.2) and (3.3) are

$$\underset{f}{\text{minimize}} \; -\int \log f(x)\,dF_n(x) \quad \text{subject to} \quad f \in L_2 \cap \mathcal{C}$$

and

$$\underset{f}{\text{minimize}} \; -\int \log f(x)\,dF_n(x) + \alpha\nu(f) \quad \text{subject to} \quad f \in L_2 \cap \mathcal{C},$$

respectively.

The functional ν, known as a *penalty functional*, is chosen to give larger values when f is "less smooth." Solutions to (3.3) are called *maximum penalized likelihood density estimators.* The parameter α is a positive real number. Typically, when $\alpha \to 0$ at some rate as $n \to \infty$, one obtains a sequence of estimators with good asymptotic properties. The literature is rich with references to work in related areas, such as regularization in Tikhonov and Arsenin [74]; maximum penalized likelihood density estimation in Good and Gaskins [21], de Montricher, Tapia, and Thompson [13], Silverman [66], and Thompson and Tapia [73]; smoothing splines in Wahba [76], [79], and

Gu [23]; nonparametric estimation and regression in Stone [70], Klonias [34], Eubank [16], Gu [24], and Härdle [27]; and inverse problems in O'Sullivan [50] and Cox [9] .

We can apply the method of penalization to problem (3.1) and obtain a related problem that does have a well-behaved unique solution. We now formulate the penalized problem.

Let $X_1, \ldots, X_n$ be i.i.d. random variables defined on a bounded domain $I \subseteq \mathbb{R}$, with c.d.f. F_o and p.d.f. $f_o = F_o'$. The function h is a smooth p.d.f. on I. Denote by $\mathcal{D}$ a linear differential operator of order $p \geqslant 1$, with no constant term, defined on a suitable domain with appropriate boundary conditions. We refer to the positive real constant α as the *smoothing parameter.* Using the penalty functional

$$\nu(f) = \frac{1}{2} \int \frac{(\mathcal{D}f)^2}{h}, \tag{3.4}$$

we define the penalized AR density estimation functional to be

$$J_{n,h}(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2} \int \frac{f^2}{h} + \alpha \nu(f). \tag{3.5}$$

The corresponding single-step AR density estimation problem is then

$$\underset{f}{\text{minimize}}\ J_{n,h}(f) \ \text{ subject to } \ f \in \mathcal{H}_p \cap \mathcal{C}, \tag{3.6}$$

and its solution $\hat{f} = \hat{f}_n$ satisfies

$$J_{n,h}(\hat{f}) = \inf \{ J_{n,h}(f) : f \in \mathcal{H}_p \cap \mathcal{C} \}. \tag{3.7}$$

For fixed n and $\hat{f}_{n,0}$, the recursive sequence of AR estimators ($\hat{f}_{n,0}$, $\hat{f}_{n,1}$, $\hat{f}_{n,2}$, $\ldots$) is characterized by

$$J_{n,\hat{f}_{n,i-1}}(\hat{f}_{n,i}) = \inf \left\{ J_{n,\hat{f}_{n,i-1}}(f) : f \in \mathcal{H}_p \cap \mathcal{C} \right\}. \tag{3.8}$$

In order to establish the existence and uniqueness of the solution of problem (3.6), we exploit the inner-product structure implicit in the form of the

penalized objective functional (3.5). Specifically, the deterministic part of $J_{n,h}(f)$ is a quadratic form that arises from a Sobolev space norm. Note also that these are the penalty functionals that generate smoothing splines.

In our setting, the Sobolev spaces $\mathcal{H}_p$ are Hilbert spaces of functions that have (Lebesgue) square-integrable p^{th} derivative.

$$\mathcal{H}_p = \{f : f^{(p)} \in L_2 \cap D\}, \tag{3.9}$$

where the domain D formally incorporates both the function support and also the boundary conditions appropriate to the application. Since domains and boundary conditions are application-specific, we generally suppress their explicit formulation and simply remember that they are a required element of the problem statement.

Some basic references for Sobolev spaces and optimization are Adams [1], Atteia [3], Kufner [35], and Luenberger [46]. However, we only need a few facts about Sobolev spaces and a theorem of Thompson and Tapia in order to give the existence and uniqueness statement. The standard Sobolev inner product is given by

$$\|f\|_p^2 = \sum_{i=0}^{p} \int \left|f^{(i)}\right|^2.$$

In our applications, a natural norm for $\mathcal{H}_p$ is given by

$$\|f\|_{p,\mathcal{D}}^2 = \int f^2 + \int (\mathcal{D}f)^2.$$

Weighted Sobolev spaces incorporate a weight function into the integral definition of the inner product, e.g.,

$$\|f\|_{p,w}^2 = \sum_{i=0}^{p} \int \left|f^{(i)}(t)\right|^2 w(t)\, dt$$

or

$$\|f\|_{p,\mathcal{D},w}^2 = \int f^2(t)\, w(t)\, dt + \int \left([\mathcal{D}f](t)\right)^2 w(t)\, dt.$$

Recall that two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on a space $\mathcal{S}$ are equivalent if there are positive real constants a and b such that

$$a\|x\|_1 \leqslant \|x\|_2 \leqslant b\|x\|_1, \quad \forall x \in \mathcal{S},$$

in which case the two norms in question give rise to the same topology, or notion of convergence, on $\mathcal{S}$.

For certain classes of weights w, the corresponding weighted spaces coincide with the unweighted spaces. This occurs, for example, if there exist δ and Δ with $0 < \delta \leqslant w(t) \leqslant \Delta < \infty$ for all t, in which case it is also obviously true that the weighted and unweighted norms are equivalent. Unweighted spaces are of course weighted with $w \equiv 1$. These various norms are related in this context by the following useful fact.

Lemma 3.2. *Let*

$$\mathcal{H}_p = \left\{ f : f^{(p)} \in L_2(I) \right\},$$

where $I \subseteq \mathbb{R}$ is bounded. Suppose there exist δ and Δ such that $w : I \to \mathbb{R}$ satisfies $0 < \delta \leqslant w(t) \leqslant \Delta < \infty$ for all $t \in I$. Suppose $\mathcal{D}$ is a linear differential operator of order $p \geqslant 1$ with no constant term. Then the following quadratic forms engender equivalent norms on $\mathcal{H}_p$:

$$\|f\|_{p,1}^2 = \sum_{i=0}^{p} \int \left| f^{(i)} \right|^2,$$

$$\|f\|_{p,p}^2 = \int f^2 + \int \left| f^{(p)} \right|^2, \quad \textit{and}$$

$$\|f\|_{p,\mathcal{D}}^2 = \int f^2 + \int |\mathcal{D} f|^2.$$

To establish the existence and uniqueness of the solution to problem (3.6), we use an optimization theorem of Thompson and Tapia [73].

Theorem 3.3. *Let $\mathcal{H}$ be a Hilbert space, let $\mathcal{C}$ be a closed convex subset of $\mathcal{H}$, and let the functional $J : \mathcal{H} \to \mathbb{R}$ be continuous in $\mathcal{C}$ and twice Gâteaux differentiable in $\mathcal{C}$ with second Gâteaux derivative uniformly positive definite in $\mathcal{C}$. Then the problem*

$$\underset{f}{\text{minimize}} \; J(f) \quad \textit{subject to} \quad f \in \mathcal{C}$$

has a unique solution in $\mathcal{C}$.

Using Lemma 3.2, we easily verify that problem (3.6) satisfies the conditions of Theorem 3.3. This implies:

Corollary 3.4. *Problem* (3.6) *has a unique solution.*

Refer to the discussion immediately following Theorem 3.1 and note that the functions $g_m = G'_m$ constructed in the proof of that theorem are discontinuous. Hence, they do not lie in $\mathcal{H}_p$ if $p \geqslant 1$. Of course, no sequence g_m in $\mathcal{H}_p$ has the property that the *penalized* objective functional $J(g_m) \to -\infty$ as $m \to \infty$.

On the other hand, if the penalization operator $\mathcal{D}$ for problem (3.6) is of order p, the problem has a unique solution with at least $p-1$ continuous derivatives. Thus, one may choose the penalization order based on which function (derivative of F_o) one wishes to estimate.

As in the parametric case, we have a theorem characterizing the limit of the estimator sequence.

Theorem 3.5. *With n fixed, if the estimator sequence associated with problem* (3.8) *converges, its limit is a maximum penalized likelihood estimator.*

The corresponding penalty functional has first Gâteaux derivative

$$\nu'(f)(r) = \int \frac{\mathcal{D}f\,\mathcal{D}r}{f}.$$

This functional itself does not seem to have a closed-form representation, although it closely resembles the Good and Gaskins [21] penalty.

3.1.1 Representation of the Density Estimator

In order to understand the properties of the penalized AR estimator, we characterize it as the solution of a certain differential equation. The solution can be represented as a generalized kernel density estimator by invoking the superposition principle for differential equations. Some details about properties of the eigenvalues and eigenfunctions of the differential equation are also useful.

Let $X_1, ..., X_n$ be i.i.d. random variables with c.d.f. F_o and p.d.f. f_o. Let $\mathcal{D}$ be a linear differential operator of order $p \geqslant 1$, with no constant term, defined on a domain with appropriate boundary conditions. For fixed h we obtain the AR density estimator as the solution $\hat{f}$ of the problem

$$\underset{f}{\text{minimize}}\ J(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2}\int \frac{f^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}f)^2}{h} \tag{3.10}$$

subject to $f \geqslant 0$ and $\int f = 1$.

The standard inner product on L_2 is $\langle x, y\rangle = \int xy$, and the corresponding square norm is $\|x\|^2 = \langle x, x\rangle$. With the weight function $w(t) = 1/h(t)$, we can identify the Hilbert space $L_{2,w}$, which has inner product $\langle x, y\rangle_w = \int xyw$. Note that $\langle x, y\rangle_w = \langle x, wy\rangle$. Let $\delta_s(t)$ denote the unit point measure at s. The empirical point measure is then $f_n = F_n' = n^{-1}\sum_{i=1}^n \delta_{X_i}$, so that $F_n(t) = \int_0^t f_n(z)\, dz = \int_0^t F_n'(z)\, dz = \int_0^t dF_n(z)$. We can now write the objective functional of (3.10) as

$$J(f) = -\langle f, f_n\rangle_w + \tfrac{1}{2}\|f\|_w^2 + \tfrac{\alpha}{2}\|\mathcal{D}f\|_w^2.$$

To describe the solution of problem (3.10), we introduce adjoint operators in Hilbert space. The L_2 adjoint $\mathcal{D}^*$ of $\mathcal{D}$ is characterized by $\langle \mathcal{D}x, y\rangle = \langle x, \mathcal{D}^*y\rangle$, with x in the domain of $\mathcal{D}$ and y in the domain of $\mathcal{D}^*$. With respect to $L_{2,w}$, the adjoint $\mathcal{D}^+$ of $\mathcal{D}$ satisfies $\langle \mathcal{D}x, y\rangle_w = \langle x, \mathcal{D}^+y\rangle_w$. The adjoints $\mathcal{D}^*$ and $\mathcal{D}^+$ are related by

$$\langle \mathcal{D}x, y\rangle_w = \langle \mathcal{D}x, wy\rangle^* = \langle x, \mathcal{D}^*wy\rangle = \left\langle x, \tfrac{1}{w}\mathcal{D}^*wy\right\rangle_w = \left\langle x, \mathcal{D}^+y\right\rangle_w.$$

Thus, we have $\mathcal{D}^+ = \frac{1}{w}\mathcal{D}^*w$. Define the operator $\mathcal{T}_w$ by

$$\mathcal{T}_w = \mathcal{D}^+\mathcal{D} = \frac{1}{w}\mathcal{D}^*w\mathcal{D},$$

where $\mathcal{T}_w$ is taken to be self-adjoint, so that formally

$$\langle \mathcal{D}x, \mathcal{D}y\rangle_w = \langle x, \mathcal{T}_w y\rangle_w = \langle \mathcal{T}_w x, y\rangle_w. \tag{3.11}$$

In what follows, we assume that (3.11) always holds.

At this point, it is beneficial to digress and look at a few specific differential operators and their adjoints. In fact, these are the operators $\mathcal{D}$ used in most of the numerical calculations in chapter 4. We set $w \equiv 1$ and $\mathcal{T} = \mathcal{T}_w = \mathcal{D}^*\mathcal{D}$ for these two examples.

First, consider $\mathcal{D}z = z'$ on the domain $[0,1]$ with the boundary conditions $z(0) = z(1) = 0$. Then we have

$$\langle \mathcal{D}x, y \rangle = \int_0^1 x'y = xy\Big|_0^1 - \int_0^1 xy' = \langle x, \mathcal{D}^*y \rangle ,$$

so that $\mathcal{D}^*z = -z'$. Note, however, that functions z in the domain of $\mathcal{D}^*$ are not necessarily subject to $z(0) = z(1) = 0$. Now observe that with x and y in the domain of $\mathcal{D}^*\mathcal{D}$, we have

$$\begin{aligned} \langle \mathcal{D}x, \mathcal{D}y \rangle = \int_0^1 x'y' = xy'\Big|_0^1 - \int_0^1 xy'' &= \langle x, \mathcal{T}y \rangle \\ = x'y\Big|_0^1 - \int_0^1 x''y &= \langle \mathcal{T}x, y \rangle , \end{aligned}$$

and (3.11) is satisfied.

Next, let $\mathcal{D}z = z''$ on $[0,1]$ with the boundary conditions $z(0) = z(1) = 0$ and $z'(0) = z'(1) = 0$. In this case, we have

$$\langle \mathcal{D}x, y \rangle = \int_0^1 x''y = x'y\Big|_0^1 - \int_0^1 x'y' = -xy'\Big|_0^1 + \int_0^1 xy'' = \langle x, \mathcal{D}^*y \rangle ,$$

so $\mathcal{D}^*z = z''$. Again, the domain of $\mathcal{D}^*$ is not subject to the boundary conditions imposed upon the domain of $\mathcal{D}$. With x and y in the domain of $\mathcal{D}^*\mathcal{D}$, we obtain

$$\begin{aligned} \langle \mathcal{D}x, \mathcal{D}y \rangle = \int_0^1 x''y'' = x'y''\Big|_0^1 - \int_0^1 x'y^{(3)} = -xy^{(3)}\Big|_0^1 + \int_0^1 xy^{(4)} &= \langle x, \mathcal{T}y \rangle \\ = x''y'\Big|_0^1 - \int_0^1 x^{(3)}y' = -x^{(3)}y\Big|_0^1 + \int_0^1 x^{(4)}y &= \langle \mathcal{T}x, y \rangle , \end{aligned}$$

and (3.11) holds here also.

To continue with the discussion, note that the penalty functional can now be written as

$$\|\mathcal{D}x\|_w^2 = \langle x, \mathcal{T}_w x \rangle_w .$$

Let $\mathcal{I}$ denote the identity transformation, and define the operators

$$\mathcal{Q}_{\alpha,w} = \mathcal{I} + \alpha \mathcal{T}_w \quad \text{and} \quad \mathcal{R}_{\alpha,w} = \mathcal{Q}_{\alpha,w}^{-1}.$$

Then we have

$$\|f\|_w^2 + \alpha \|\mathcal{D}f\|_w^2 = \langle f, \mathcal{Q}_{\alpha,w} f \rangle , \tag{3.12}$$

where $\mathcal{Q}_{\alpha,w}$ is a self-adjoint operator. The AR objective functional and its first and second Gâteaux derivatives now become

$$\begin{aligned} J(f) &= -\langle f, f_n \rangle_w + \tfrac{1}{2} \langle f, \mathcal{Q}_{\alpha,w} f \rangle_w , \\ J'(f)(r) &= -\langle r, f_n \rangle_w + \langle r, \mathcal{Q}_{\alpha,w} f \rangle_w , \quad \text{and} \\ J''(f)(r,s) &= \langle r, \mathcal{Q}_{\alpha,w} s \rangle_w . \end{aligned}$$

To accomplish the minimization of $J(f)$, we let $J'(f)(r) = 0$ for all r. Thus, we obtain the weak differential equation representation of the AR estimator as the solution of

$$\mathcal{Q}_{\alpha,w} f = f_n. \tag{3.13}$$

In terms of the inverse operator, the solution can be written as

$$\hat{f} = \mathcal{R}_{\alpha,w} f_n. \tag{3.14}$$

To develop the generalized kernel representation of the AR estimator, we introduce the kernel $Z_{\alpha,w,s}$, which satisfies the equation

$$\mathcal{Q}_{\alpha,w} Z_{\alpha,w,s} = \delta_s. \tag{3.15}$$

Since $f_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$, we can use (3.15) to write f_n as

$$f_n = \frac{1}{n} \sum_{i=1}^n \mathcal{Q}_{\alpha,w} Z_{\alpha,w,X_i} = \mathcal{Q}_{\alpha,w} \left[\frac{1}{n} \sum_{i=1}^n Z_{\alpha,w,X_i} \right].$$

Then, by (3.13), we have

$$\mathcal{Q}_{\alpha,w} f = \mathcal{Q}_{\alpha,w} \left[\frac{1}{n} \sum_{i=1}^{n} Z_{\alpha,w,X_i} \right].$$

This yields the generalized kernel representation of the AR estimator—namely,

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} Z_{\alpha,w,X_i}(t) = \int Z_{\alpha,w,s}(t)\, dF_n(s). \tag{3.16}$$

3.1.2 Special Cases

Now we take a brief look at two particular examples. These are interesting in one respect because we can perform the calculations and exhibit closed-form solutions. However, and more importantly, we see in these cases that the ARE is in fact a kernel density estimator with a familiar kernel. So, the properties (including asymptotic theory) of this ARE are already well investigated.

3.1.2.1 Boundary-Corrected Kernel Density Estimator. In this example, we use the first-order differential penalty operator $\mathcal{D}x = x'$ on $[0,1]$ with boundary conditions $x'(0) = x'(1) = 0$. The reference distribution is uniform, so $H(t) = t$ and $H'(t) = h(t) = w(t) = 1$. For notational convenience we use $\beta = \alpha^{-1/2}$ as the smoothing parameter, and we write $Z_{\beta,s}$ in place of $Z_{\beta,w,s}$. The adjoint operator is $\mathcal{D}^* x = -x'$ with boundary conditions $x(0) = x(1) = 0$, and the quadratic form operator is given by $\mathcal{Q}_\beta f = f - \beta^{-2} f''$. Then the kernel $Z_{\beta,s}$ satisfies the weak differential equation

$$Z_{\beta,s}(t) - \beta^{-2} Z''_{\beta,s}(t) = \delta_s(t) \tag{3.17}$$

subject to the boundary conditions $Z'_{\beta,s}(0) = Z'_{\beta,s}(1) = 0$. The solution is seen to be

$$Z_{\beta,s}(t) = \beta \operatorname{csch} \beta \cosh[\beta(s \wedge t)] \cosh[\beta(1 - s \vee t)]$$

for $(s,t) \in [0,1] \times [0,1]$. We verify that this is the solution. Let $Z_{\beta,s}(t) = U'(t)$, where

$$U(t) - \beta^{-2}U''(t) = \begin{cases} 0, & t \leqslant s \\ 1, & t > s \end{cases}$$

subject to $U(0) = U''(0) = U''(1) = 0$. With

$$U(t) = \begin{cases} \operatorname{csch}\beta \cosh[\beta(1-s)] \sinh(\beta t), & t \leqslant s \\ 1 - \operatorname{csch}\beta \cosh(\beta s) \sinh[\beta(1-t)], & t > s, \end{cases}$$

$$Z_{\beta,s}(t) = U'(t) = \begin{cases} \beta \operatorname{csch}\beta \cosh[\beta(1-s)] \cosh(\beta t), & t \leqslant s \\ \beta \operatorname{csch}\beta \cosh(\beta s) \cosh[\beta(1-t)], & t > s, \end{cases} \text{ and}$$

$$Z'_{\beta,s}(t) = U''(t) = \begin{cases} \beta^2 \operatorname{csch}\beta \cosh[\beta(1-s)] \sinh(\beta t), & t \leqslant s \\ -\beta^2 \operatorname{csch}\beta \cosh(\beta s) \sinh[\beta(1-t)], & t > s, \end{cases}$$

we see that the equation and boundary conditions are satisfied. We can use the standard identities

$$\begin{aligned} \sinh(x+y) &= \sinh x \cosh y + \cosh x \sinh y, \\ \cosh(x+y) &= \cosh x \cosh y + \sinh x \sinh y, \text{ and} \\ 1 &= \cosh^2 x - \sinh^2 x \end{aligned}$$

to see that U is continuous at s. Thus, in this case, the ARE is a boundary-corrected kernel density estimator with a bilateral exponential kernel.

3.1.2.2 Kernel Density Estimator. The example of the previous section can be extended to the real line as follows. We use the same differential penalty operator $\mathcal{D}x = x'$ on the interval $[-M, M]$. The boundary conditions for $\mathcal{D}$ are $x'(-M) = x'(M) = 0$, and the solution of (3.17) is then

$$Z_{\beta,s}(t) = \beta \operatorname{csch}(2\beta M) \cosh[\beta(s \wedge t + M)] \cosh[\beta(M - s \vee t)]$$

on $[-M, M]$. Equivalently, we can write

$$Z_{\beta,s}(t) = \begin{cases} \beta \operatorname{csch}(2\beta M) \cosh[\beta(t+M)] \cosh[\beta(M-s)], & -M \leqslant t \leqslant s \\ \beta \operatorname{csch}(2\beta M) \cosh[\beta(s+M)] \cosh[\beta(M-t)], & s < t \leqslant M. \end{cases}$$

Since $\sinh x \sim \cosh x \sim \frac{1}{2}e^x$ as $x \to \infty$, we have for $t \leqslant s$

$$Z_{\beta,s}(t) \sim \frac{\beta \cdot \frac{1}{2}e^{\beta(t+M)} \cdot \frac{1}{2}e^{\beta(M-s)}}{\frac{1}{2}e^{2\beta M}} = \frac{\beta}{2}e^{\beta(t-s)}$$

and for $t > s$

$$Z_{\beta,s}(t) \sim \frac{\beta \cdot \frac{1}{2}e^{\beta(s+M)} \cdot \frac{1}{2}e^{\beta(M-t)}}{\frac{1}{2}e^{2\beta M}} = \frac{\beta}{2}e^{\beta(s-t)}.$$

Putting this together gives

$$Z_{\beta,s}(t) \sim K_{\beta,s}(t) \equiv \frac{\beta}{2}e^{-\beta|s-t|},$$

and as $M \to \infty$ the AR density estimator $\hat{f}(t) = \int_{-M}^{M} Z_{\beta,s}(t)\, dF_n(s)$ satisfies

$$\hat{f}(t) \sim \tilde{f}(t) \equiv \int_{-\infty}^{\infty} K_{\beta,s}(t)\, dF_n(s).$$

Of course, $\tilde{f}(t)$ is the convolution kernel density estimator with a bilateral exponential kernel. In fact, since

$$\int_{-\infty}^{\infty} |Z_{\beta,s}(t) - K_{\beta,s}(t)|\, dt = 2e^{-\beta M}\cosh \beta s,$$

we can establish that

$$\|\hat{f}(t) - \tilde{f}(t)\|_{L_1} \leqslant 2e^{-\beta M}\sum_{i=1}^{n}\cosh \beta X_i.$$

Therefore, this AR estimator on $[-M, M]$ converges in L_1 to a kernel density estimator as $M \to \infty$.

3.1.3 Consistency and Rates of Convergence

Consistency results and convergence rates are expressed in terms of distance measures on the space of functions containing the true parameter, f_o, and the estimates. To obtain results, one must typically use a sequence of smoothing

parameters α_n satisfying $\alpha_n \to 0$ as $n \to \infty$, where n is the sample size. We make these dependencies explicit by writing the solution of (3.7) as

$$\hat{f} = \hat{f}_{n,\alpha_n,w}. \tag{3.18}$$

Then one considers the convergence (in some sense) of $\hat{f}_{n,\alpha_n,w}$ to f_o as $n \to \infty$.

On the other hand, for fixed n and sufficiently smooth p.d.f. $\hat{f}_{n,0}$, we have the recursive AR estimator sequence $(\hat{f}_{n,0}, \hat{f}_{n,1}, \hat{f}_{n,1}, \dots)$ characterized by (3.8)—namely,

$$\hat{f}_{n,i+1} = \hat{f}_{n,\alpha_n,w_i}, \quad \text{where} \quad w_i = 1/\hat{f}_{n,i}. \tag{3.19}$$

In the following subsections we consider the properties of a single-stage AR estimator (3.18), which means that w is taken to be fixed. We are not concerned at the present time with the properties of the recursive ARE sequence (3.19).

We can approach these problems on a case-by-case basis, considering individual (differential operator and boundary condition) configurations as in section 3.1.2. Alternatively, we can adopt a more general approach. At least two essentially different general approaches to similar problems appear in the literature. One is Bosq and Lecoutre's [6] probability-theoretic analysis of generalized kernel density estimators. Another is the spectral analysis method used by Silverman, Cox, O'Sullivan, and Wahba. (See the references on page 53.) We now apply each of these techniques in an attempt to obtain general results about the consistency and rates of convergence of the single-step AR density estimator.

3.1.3.1 Generalized Kernel Analysis. For an AR estimator $\hat{f}_{n,\alpha_n,w}$ of f_o with domain I, we define the pointwise distance measure

$$D_n(t) = \left| \hat{f}_{n,\alpha_n,w}(t) - f_o(t) \right|, \quad t \in I.$$

For $G \subseteq I$, we use the restricted global distance measure

$$d_G(g, f) = \sup_{t \in G} |g(t) - f(t)|.$$

Now consider the AR density estimator constructed with a fixed (initial guess) weight function w. From section 3.1.1, we have the generalized kernel density estimator representation

$$\hat{f}_{n,\alpha_n,w}(t) = \frac{1}{n}\sum_{i=1}^{n} Z_{\alpha_n,w,X_i}(t) = \int_I Z_{\alpha_n,w,s}(t)\, dF_n(s),$$

where the kernel $Z_{\alpha,w,s}$ satisfies the equation

$$\mathcal{Q}_{\alpha,w} Z_{\alpha,w,s} = \delta_s.$$

The generalized kernel density estimator has been analyzed by Bosq and Lecoutre [6]. We apply their theorem in the case of a fixed weight w and state the result here.

Let $I = [0,1]$. We consider the measure space $(I, \mathcal{B}, \lambda)$, where λ is the Lebesgue measure on I and $\mathcal{B}$ is the Borel σ-algebra. The space $\mathcal{F}$ is the functional domain of the AR estimation problem. For $G \subseteq I$, define the space $\mathcal{F}_G$ by

$$\mathcal{F}_G = \left\{ f : f \in \mathcal{F}; \lim_{\alpha \to 0} \sup_{t \in G} \left| \left[\left(\mathcal{R}_{\alpha,w} - \mathcal{I} \right) f \right](t) \right| = 0 \right\}.$$

The intent is that $\mathcal{F}_G$ consists of functions f that are sufficiently well-behaved on G so that $\mathcal{R}_{\alpha,w} f \to f$ in the manner indicated as $\alpha \to 0$.

The five conditions that follow are part of the necessary and sufficient criteria for Bosq and Lecoutre's results on arbitrary kernels. Our kernels arise as solutions of certain linear differential equations, which should imply that the first four conditions hold. The conditions can be verified on a case-by-case basis for individual combinations of differential operator, boundary conditions, and domain. Loosely speaking, conditions (1) and (2) are bounds on the supremum and L_2 norm of Z, respectively, and conditions (3) and (4) relate to the continuity of $Z_{\alpha,w,x}(t)$ in x and t, respectively. Condition (5) involves the metric structure of I and G, and holds when G is an interval. The conditions follow.

Suppose that there is a positive constant β, and:

(1) There is a bounded function $A : G \to \mathbb{R}$, and for each $t \in G$ there is an α_o such that if $\alpha < \alpha_o$ then

$$\sup_{x \in I} |Z_{\alpha,w,x}(t)| \leqslant \alpha^{-\beta} A(t).$$

(2) There is a bounded function $B : G \to \mathbb{R}$, and for each $t \in G$ there is an α_o such that if $\alpha < \alpha_o$ then

$$\int_I Z^2_{\alpha,w,x}(t)\, dx \leqslant \alpha^{-\beta} B(t).$$

(3) There are a function $g_o \in \mathcal{F}_G$, a point $t_o \in G$, and a family $\{\mathcal{B}_\alpha : \alpha > 0\}$ of Borel subsets of I such that

$$\begin{aligned} &g_o(t_o) > 0, \\ &\lambda(\mathcal{B}_\alpha) = k\alpha^\beta \quad \text{for} \quad k > 0, \quad \text{and} \\ &\int_{\mathcal{B}_\alpha} Z_{\alpha,w,x}(t_o) g_o(x)\, dx \geqslant \delta \quad \text{for} \quad \delta > 0. \end{aligned}$$

(4) $(I, \mathcal{B})$ is a metric space with Borel σ-algebra $\mathcal{B}$, and $Z_{\alpha,w,x}$ satisfies the Lipschitz condition

$$\sup_{x \in I} |Z_{\alpha,w,x}(t') - Z_{\alpha,w,x}(t)| \leqslant C\alpha^{-m} |t - t'|^\gamma$$

for some $C > 0$, $m > 0$, and $\gamma > 0$.

(5) $(I, \mathcal{B})$ is a metric space with Borel σ-algebra $\mathcal{B}$, and $G \in \mathcal{B}$ is precompact (i.e., has compact closure). $0 < \lambda(G) < \infty$, and there is an $h > 0$ such that for all ε small enough, G can be covered by $[\varepsilon^{-h}]$ balls of radius $\leqslant \varepsilon$.

Next is a condition on the convergence of the smoothing parameter sequence. The sequence a_n is called *asymptotically concave* if there are a concave function g, positive constants c_1 and c_2, and an integer n_o such that

$$c_1 g(n) \leqslant a_n \leqslant c_2 g(n), \quad \forall n > n_o.$$

The mode of convergence used in Theorem 3.6 is *almost complete convergence*, denoted $X_n \xrightarrow{a.co.} X$ and defined by

$$X_n \xrightarrow{a.co.} X \iff \forall \varepsilon > 0, \sum_{n=1}^{\infty} \Pr[d(X_n, X) \geqslant \varepsilon] < \infty.$$

Almost complete convergence implies *almost sure convergence*, denoted $X_n \xrightarrow{a.s.} X$ and defined by

$$X_n \xrightarrow{a.s.} X \iff \Pr[d(X_n, X) \to 0] = 1.$$

In turn, almost sure convergence implies *convergence in probability*, denoted $X_n \xrightarrow{p} X$ and defined by

$$X_n \xrightarrow{p} X \iff \forall \varepsilon > 0, \Pr[d(X_n, X) \geqslant \varepsilon] \to 0.$$

Finally, we state the results of Bosq and Lecoutre. The theorem guarantees strong uniform consistency of the generalized kernel density estimator. The corollary establishes the rate of convergence.

Theorem 3.6. *If the preceding conditions (1) through (5) hold and $\alpha_n^{-\beta}$ is asymptotically concave, then the following conditions are equivalent.*

(1) $n^{-1}\alpha_n^{-\beta} \log n \to 0$.

(2) $D_n(t) \xrightarrow{a.co.} 0$, $\mathrm{E}\, D_n(t) \to 0$; $t \in G$, $f_o \in \mathcal{F}_G$.

(3) $d_G(\hat{f}_{n,\alpha_n,w}, f_o) \xrightarrow{a.co.} 0$, $\mathrm{E}\, d_G(\hat{f}_{n,\alpha_n,w}, f_o) \to 0$; $f_o \in \mathcal{F}_G$.

Corollary 3.7. *Under the preceding conditions, there is a $\delta > 0$ such that for all n large enough and for all $\varepsilon > 0$*

$$\Pr[d_G(\hat{f}_{n,\alpha_n,w}, f_o) \geqslant \varepsilon] < 2\exp(-\delta\varepsilon^2 n\alpha_n^{\beta}).$$

As a consequence of the corollary, if we choose

$$\varepsilon_n = \lambda \cdot (\log n)^{1/2} n^{-1/2} \alpha_n^{-\beta/2},$$

we then obtain

$$\Pr\left[(\log n)^{-1/2} n^{1/2} \alpha_n^{\beta/2}\, d_G(\hat{f}_{n,\alpha_n,w}, f_o) \geqslant \lambda\right] < 2\exp(-\delta\lambda^2 \log n) = 2n^{-\delta\lambda^2}$$

for any $\lambda > 0$. Then the corresponding rate of convergence in probability is

$$(\log n)^{-1/2} n^{1/2} \alpha_n^{\beta/2}\, d_G(\hat{f}_{n,\alpha_n,w}, f_o) \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.$$

We can show that $\beta = (2p)^{-1}$, where p is the order of the differential penalization operator. The convergence conditions are then

$$\alpha_n \to 0 \quad \text{and} \quad \alpha_n \left[\frac{n}{\log n}\right]^{2p} \to \infty \quad \text{as} \quad n \to \infty.$$

Thus, the rate for convergence in probability becomes

$$(\log n)^{-1/2} n^{1/2} \alpha_n^{1/4p}\, d_G(\hat{f}_{n,\alpha_n,w}, f_o) \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty,$$

and choosing $\alpha_n = n^{-2\gamma p}$ for some γ with $0 < \gamma < 1$ results in

$$(\log n)^{-1/2} n^{(1-\gamma)/2}\, d_G(\hat{f}_{n,\alpha_n,w}, f_o) \xrightarrow{p} 0 \quad \text{as} \quad n \to \infty.$$

3.1.3.2 Spectral Analysis. Here, f_o is the true parameter value. We analyze the estimator $\hat{f} = \mathcal{R}_{\alpha,w} f_n$ by characterizing the error as

$$\hat{f} - f_o = (\hat{f} - \mathcal{R}_{\alpha,w} f_o) + (\mathcal{R}_{\alpha,w} f_o - f_o),$$

interpreting $\mathcal{R}_{\alpha,w} f_o$ as the asymptotic "infinite sample size" solution of the estimation problem. In any norm $\|\cdot\|$, we can bound the error of the estimator by means of

$$\tfrac{1}{2}\|\hat{f} - f_o\|^2 \leqslant \|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2 + \|\mathcal{R}_{\alpha,w} f_o - f_o\|^2 = T_1 + T_2,$$

in which

$$T_1 = \|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2 - \mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2 \quad \text{and}$$
$$T_2 = \mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2 + \|\mathcal{R}_{\alpha,w} f_o - f_o\|^2.$$

Since $\mathrm{E}\, T_1 = 0$, it follows that $\mathrm{E}\,\|\hat{f} - f_o\|^2 \leqslant 2T_2$. The "mean" error T_2 is the sum of a "variance" term $\mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2$ and a "bias" term $\|\mathcal{R}_{\alpha,w} f_o - f_o\|^2$. It is shown in Theorem 3.8 that $\mathrm{E}\,\hat{f} = \mathcal{R}_{\alpha,w} f_o$. We now characterize the behavior of T_2 to obtain rates of convergence for $\mathrm{E}\,\|\hat{f} - f_o\|^2$.

To that end, let λ_i and ψ_i be the eigenvalues and $L_{2,w}$-orthonormal eigenfunctions of $\mathcal{Q}_{1,w}$, so that

$$\mathcal{Q}_{1,w}\psi_i = \lambda_i^{-1}\psi_i, \quad \langle \psi_i, \mathcal{Q}_{1,w}\psi_j \rangle_w = \lambda_i^{-1}\delta_{ij}, \quad \text{and} \quad \langle \psi_i, \psi_j \rangle_w = \delta_{ij}.$$

Recall that $\mathcal{T}_w = \frac{1}{w}\mathcal{D}^* w \mathcal{D}$, where $\mathcal{D}$ is a linear differential operator of order p with no constant term. Note that we are using the operator $\mathcal{Q}_{1,w} = \mathcal{I} + \mathcal{T}_w$, and that

$$\mathcal{Q}_{\alpha,w} = (1-\alpha)\mathcal{I} + \alpha\mathcal{Q}_{1,w} = \mathcal{I} + \alpha(\mathcal{Q}_{1,w} - \mathcal{I}).$$

We use $\mathcal{Q}_{1,w}$ in the analysis to isolate the smoothing parameter α. By equation (3.12), we have

$$\langle f, \mathcal{Q}_{1,w} f \rangle_w = \|f\|_w^2 + \|\mathcal{D}f\|_w^2$$

for all f. Therefore, $\mathcal{Q}_{1,w}$ is a positive operator and all eigenvalues are positive. With $\psi_1 \equiv$ constant, we see that $\mathcal{Q}_{1,w}\psi_1 = \psi_1$, and so the constant function ψ_1 is an eigenfunction with eigenvalue $\lambda_1 = 1$. Furthermore, since we have on one hand

$$\langle \psi_i, \mathcal{Q}_{1,w}\psi_i \rangle_w = \|\psi_i\|_w^2 + \|\mathcal{D}\psi_i\|_w^2 = 1 + \|\mathcal{D}\psi_i\|_w^2$$

and on the other hand

$$\langle \psi_i, \mathcal{Q}_{1,w}\psi_i \rangle_w = \langle \psi_i, \lambda_i^{-1}\psi_i \rangle_w = \lambda_i^{-1} \langle \psi_i, \psi_i \rangle_w = \lambda_i^{-1}\|\psi_i\|_w^2 = \lambda_i^{-1},$$

it follows that

$$\lambda_i = \frac{1}{1 + \|\mathcal{D}\psi_i\|_w^2}.$$

Therefore, the eigenvalues are all positive, and they satisfy

$$1 = \lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots.$$

Also, since $\mathcal{T}_w$ is a linear differential operator of order $2p$, one can show that there are constants a and b with

$$0 < a \leqslant \lambda_i \, i^{2p} \leqslant b < \infty,$$

so that the eigenvalues λ_i decay like i^{-2p}. For reference, see Riesz and Nagy [58], p. 238, and Silverman [66], Lemma 5.1.

The ψ_i are also $L_{2,w}$-orthonormal eigenfunctions for $\mathcal{Q}_{\alpha,w}$ and $\mathcal{R}_{\alpha,w}$. These operators act on the eigenfunctions as follows.

$$\mathcal{Q}_{\alpha,w}\psi_i = \gamma_i^{-1}\psi_i \quad \text{and} \quad \mathcal{R}_{\alpha,w}\psi_i = \gamma_i\psi_i,$$

where

$$\gamma_i^{-1} = (1-\alpha) + \alpha\lambda_i^{-1} = 1 + \alpha(\lambda_i^{-1} - 1).$$

General eigenfunction expansions are then

$$x = \sum_{i=1}^{\infty} \langle x, \psi_i \rangle_w \, \psi_i, \qquad \mathcal{Q}_{1,w}x = \sum_{i=1}^{\infty} \lambda_i^{-1} \langle x, \psi_i \rangle_w \, \psi_i,$$

$$\mathcal{Q}_{\alpha,w}x = \sum_{i=1}^{\infty} \gamma_i^{-1} \langle x, \psi_i \rangle_w \, \psi_i, \quad \text{and} \quad \mathcal{R}_{\alpha,w}x = \sum_{i=1}^{\infty} \gamma_i \langle x, \psi_i \rangle_w \, \psi_i.$$

For the analysis of AR estimators, relevant eigenfunction expansions are

$$f_o = \sum_{i=1}^{\infty} c_i\psi_i, \qquad \mathcal{R}_{\alpha,w}f_o = \sum_{i=1}^{\infty} \gamma_i c_i \psi_i, \qquad \hat{f} = \sum_{i=1}^{\infty} \gamma_i \beta_i \psi_i,$$

$$\hat{f} - \mathcal{R}_{\alpha,w}f_o = \sum_{i=1}^{\infty} \gamma_i(\beta_i - c_i)\psi_i, \quad \text{and} \quad \mathcal{R}_{\alpha,w}f_o - f_o = \sum_{i=1}^{\infty} (\gamma_i - 1)c_i\psi_i,$$

where

$$c_i = \langle f_o, \psi_i \rangle_w = \int f_o \psi_i w = \int \psi_i w \, dF_o \quad \text{and}$$

$$\beta_i = \langle f_n, \psi_i \rangle_w = \int f_n \psi_i w = \int \psi_i w \, dF_n.$$

Now, note that $\mathrm{E}\,\beta_i = c_i$. This implies $\mathrm{E}\,\hat{f} = \mathcal{R}_{\alpha,w}f_o$, which in turn establishes:

Theorem 3.8. *For any weight w, the AR density estimator $\hat{f} = \mathcal{R}_{\alpha,w} f_n$ is an unbiased estimator of $\mathcal{R}_{\alpha,w} f_o$, where f_o is the true parameter value.*

The condition $\|\Omega_{1,w} f\|_w^2 < \infty$ is equivalent to Wahba's [78] "very smooth." We refer to the stronger condition $\|\Omega_{1,w} f\|_{1,w}^2 < \infty$ as "way smooth." Convergence of the bias $\mathcal{R}_{\alpha,w} f_o - f_o$ is provided by:

Lemma 3.9. *Suppose that $\|\Omega_{1,w} f_o\|_w^2 < \infty$. Then*

$$\begin{aligned} \|\mathcal{R}_{\alpha,w} f_o - f_o\|_w^2 &= O(\alpha^2) \quad \textit{and} \\ \|\mathcal{R}_{\alpha,w} f_o - f_o\|_{1,w}^2 &= O(\alpha). \end{aligned}$$

If in addition $\|\Omega_{1,w} f_o\|_{1,w}^2 < \infty$, then

$$\|\mathcal{R}_{\alpha,w} f_o - f_o\|_{1,w}^2 = O(\alpha^2).$$

We can also provide rates of convergence for the error variance term $\hat{f} - \mathcal{R}_{\alpha,w} f_o$. Note that the rates of convergence given in Lemma 3.10 apply when the weight function has the "correct" value of $w = 1/f_o$.

Lemma 3.10. *Suppose that $w = 1/f_o$ and $\|\Omega_{1,w} f_o\|_w^2 < \infty$. Then*

$$\begin{aligned} \mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|_w^2 &= O\left(n^{-1}\alpha^{-1/2p}\right) \quad \textit{and} \\ \mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|_{1,w}^2 &= O\left(n^{-1}\alpha^{-1-1/2p}\right). \end{aligned}$$

Combining Lemma 3.9 and Lemma 3.10 establishes

Theorem 3.11. *Suppose that $w = 1/f_o$ and $\|\Omega_{1,w} f_o\|_w^2 < \infty$. Then*

$$\begin{aligned} \mathrm{E}\,\|\hat{f} - f_o\|_w^2 &= O\left(n^{-1}\alpha^{-1/2p} + \alpha^2\right) \quad \textit{and} \\ \mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 &= O\left(n^{-1}\alpha^{-1-1/2p} + \alpha\right). \end{aligned}$$

If in addition $\|\Omega_{1,w} f_o\|_{1,w}^2 < \infty$, then

$$\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 = O\left(n^{-1}\alpha^{-1-1/2p} + \alpha^2\right).$$

This result implies the following statements about rates of convergence under various conditions. In the norm $\|\cdot\|_w$, convergence is obtained for any sequence α_n with

$$\alpha_n \to 0 \quad \text{and} \quad n^{2p}\alpha_n \to \infty \quad \text{as} \quad n \to \infty.$$

In particular, with

$$\alpha_n \sim n^{-2p/(4p+1)} = n^{-1/2+1/(8p+2)},$$

the best rate of convergence provided by Theorem 3.11 is

$$\mathrm{E}\,\|\hat{f} - f_o\|_w^2 = O\left(n^{-4p/(4p+1)}\right) = O\left(n^{-1+1/(4p+1)}\right).$$

In the norm $\|\cdot\|_{1,w}$, convergence is obtained for any sequence α_n with

$$\alpha_n \to 0 \text{ and } n^{2p/(2p+1)}\alpha_n \to \infty \quad \text{as} \quad n \to \infty,$$

and when

$$\alpha_n \sim n^{-2p/(4p+1)} = n^{-1/2+1/(8p+2)}$$

the best rate of convergence provided by Theorem 3.11 is

$$\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 = O\left(n^{-2p/(4p+1)}\right) = O\left(n^{-1/2+1/(8p+2)}\right).$$

Suppose additionally that f_o is way smooth and

$$\alpha_n \sim n^{-2p/(6p+1)} = n^{-1/3+1/(18p+3)}.$$

Then the best rate of convergence provided by Theorem 3.11 is

$$\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 = O\left(n^{-4p/(6p+1)}\right) = O\left(n^{-2/3+2/(18p+3)}\right).$$

It remains to characterize the random error component $\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|^2$ in order to obtain convergence rates for $\|\hat{f} - f_o\|^2 \xrightarrow{p} 0$.

3.2 Inverse Problems

In this section, we discuss a generalization of density estimation that involves indirectly observed data. We present two examples, the deconvolution problem and Wicksell's corpuscle problem.

Let $X_1, \ldots, X_n$ be i.i.d. with unknown p.d.f. f_o, which we wish to estimate. Suppose that the X_i's are not directly observable, but that we do observe the "transformed" i.i.d. data $Z_1, \ldots, Z_n$ where the common p.d.f. of the Z_i's is

$$g_o(t) = [\mathcal{K} f_o](t)$$

for some known operator $\mathcal{K}$. Based on the data $Z_1, \ldots, Z_n$, the AR functional for estimation of g_o is

$$J(g) = -\langle g_n, g \rangle_w + \tfrac{1}{2}\|g\|_w^2.$$

The "correct" weight is $w = 1/g$, and $g_n = G_n'$ where the empirical c.d.f. G_n is based on the observable data $Z_1, \ldots, Z_n$.

Since $g = \mathcal{K} f$, we estimate f using the AR objective functional

$$J(f) = -\langle g_n, \mathcal{K} f \rangle_w + \tfrac{1}{2}\|\mathcal{K} f\|_w^2.$$

For nonparametric estimation, we penalize f and use

$$J(f) = -\langle g_n, \mathcal{K} f \rangle_w + \tfrac{1}{2}\|\mathcal{K} f\|_w^2 + \tfrac{\alpha}{2}\|\mathcal{D} f\|_w^2,$$

where the "correct" weight is $w = 1/\mathcal{K} f$. The penalty operator has the usual characteristics: $\mathcal{D}$ is a linear differential operator of order $p \geqslant 1$, with no constant term, defined on a suitable domain with appropriate boundary conditions.

Let $(\mathcal{K}' f)(r)$ denote the Gâteaux derivative of $\mathcal{K}$ at f in the direction r. Then the Gâteaux derivative of J at f in the direction r is

$$\begin{aligned} J'(f)(r) &= -\langle g_n, (\mathcal{K}' f)(r) \rangle_w + \langle \mathcal{K} f, (\mathcal{K}' f)(r) \rangle_w + \alpha \langle \mathcal{D} f, \mathcal{D} r \rangle_w \\ &= -\langle w g_n, (\mathcal{K}' f)(r) \rangle + \langle w \mathcal{K} f, (\mathcal{K}' f)(r) \rangle + \alpha \langle w \mathcal{D} f, \mathcal{D} r \rangle \\ &= -\langle (\mathcal{K}' f)^* w g_n, r \rangle + \langle (\mathcal{K}' f)^* w \mathcal{K} f, r \rangle + \alpha \langle \mathcal{D}^* w \mathcal{D} f, r \rangle, \end{aligned}$$

and the differential equation for the estimator f of f_o is

$$([(\mathcal{K}'f)^* w\mathcal{K} + \alpha\mathcal{D}^* w\mathcal{D}]\, f)\,(t) = [(\mathcal{K}'f)^* w g_n]\,(t) \tag{3.20}$$

or, more briefly,

$$[(\mathcal{K}'f)^* w\mathcal{K} + \alpha\mathcal{D}^* w\mathcal{D}]\, f = (\mathcal{K}'f)^* w g_n.$$

This applies when $\mathcal{K}$ is an arbitrary (linear or nonlinear) operator.

If $\mathcal{K}$ is a linear operator, then $\mathcal{K}'f = \mathcal{K}$. In this case, the differential equation simplifies to

$$[\mathcal{K}^* w\mathcal{K} + \alpha\mathcal{D}^* w\mathcal{D}]f = \mathcal{K}^* w g_n. \tag{3.21}$$

It is possible to conduct a spectral analysis of the linear inverse problem using the technique of "simultaneous diagonalization" for infinite-dimensional operators and, thereby, obtain general results. See Cox [9] and Cox and O'Sullivan [10] for work of this nature. We do not pursue this analysis here.

3.2.1 Deconvolution

Consider the model

$$X = Z + W$$

where the random variables have densities

$$X \sim g_o, \quad Z \sim f_o, \quad \text{and} \quad W \sim k.$$

We assume that Z and W are independent and that k is a known continuous density. The parameter of interest is the density f_o of Z. However, we observe X and not Z. The cumulative distribution of X is

$$G_o(t) = \Pr(X \leqslant t) = \Pr(Z \leqslant t - W) = \int_{-\infty}^{\infty} \int_{\infty}^{t-w} f_o(z)k(w)\, dz\, dw,$$

so X has density

$$g_o(t) = \int_{-\infty}^{\infty} f_o(t-w)k(w)\, dw = \int_{-\infty}^{\infty} k(t-x)f_o(x)\, dx.$$

This is called the *convolution* of k and f_o, and denoted

$$[k * f_o](t) = \int_{-\infty}^{\infty} k(t-x) f_o(x)\, dx.$$

So, in the context of section 3.2, we have $[\mathcal{K} f_o](t) = [k * f_o](t)$. Since $\mathcal{K}$ is a linear operator, the equation for the AR estimator $\hat{f}$ of f_o is given by (3.21).

This problem can be analyzed directly for a specific contaminating distribution k, at an intermediate level of generalization with $\mathcal{K}$ being the convolution operator for an arbitrary k, or through the general methods referenced at the end of section 3.2. We pursue no further analysis here, but in chapter 4 we do present some example calculations and a simulation study that compares the nonparametric AR deconvolution estimator to other deconvolution estimators.

3.2.2 The Corpuscle Problem

Imagine a solid medium in which spheres occur according to a homogeneous Poisson process with unknown rate λ. The sphere radius is the random variable of interest, with p.d.f. f_o, which has support $[0, R_M]$ where $R_M > 0$. The radii are not observed directly. A planar slice is taken through the medium, and we observe the resulting radii of circles that are the intersection of the plane and certain spheres. Of course, the slice misses some spheres completely and cuts the others at unknown latitudes. The p.d.f. g_o of the observable circle radii is calculated in the following manner.

Let the sphere radius R have p.d.f. $f_o I_{[0,R_M]}$. Independently of R, let the random variable Y have the uniform distribution on $(-R_M, R_M)$. So the p.d.f. of Y is

$$\frac{1}{2R_M} I_{(-R_M, R_M)},$$

and Y represents the sphere coordinate at which the slice is taken. Let the random variable δ be given by

$$\delta = \begin{cases} 1, & |Y| < R \\ 0, & |Y| \geqslant R, \end{cases}$$

so that $\delta = 1$ if and only if the sphere is sliced, in which event $D^2 + Y^2 = R^2$, where D is the circle radius. The c.d.f. of circle radius is denoted by G_o, and so

$$1 - G_o(t) = \Pr(D > t | \delta = 1) = \frac{\Pr(D > t \text{ and } \delta = 1)}{\Pr(\delta = 1)}.$$

The denominator is

$$\Pr(\delta = 1) = \int_0^{R_M} \int_{-r}^{r} \frac{1}{2R_M} f_o(r)\, dy\, dr = \frac{1}{R_M} \int_0^{R_M} r f_o(r)\, dr = \frac{1}{R_M} \mathrm{E}\, R.$$

Since $D > t$ and $\delta = 1$ together imply that both $t < R < R_M$ and $Y^2 < R^2 - t^2$, we calculate the numerator as

$$\int_t^{R_M} \int_{-\sqrt{r^2-t^2}}^{\sqrt{r^2-t^2}} \frac{1}{2R_M}\, dy\, f_o(r)\, dr = \frac{1}{R_M} \int_t^{R_M} \sqrt{r^2 - t^2} f_o(r)\, dr.$$

Then the circle radius c.d.f. is

$$G_o(t) = 1 - \frac{1}{\mathrm{E}\, R} \int_t^{R_M} \sqrt{r^2 - t^2} f_o(r)\, dr,$$

and the p.d.f. is

$$g_o(t) = \frac{1}{\mathrm{E}\, R} \int_t^{R_M} \frac{t}{\sqrt{r^2 - t^2}} f_o(r)\, dr.$$

So we can write

$$g_o = \mathcal{K} f_o,$$

where the operator is given by

$$[\mathcal{K} f](t) = \frac{t \int_t^{R_M} (r^2 - t^2)^{-1/2} f(r)\, dr}{\int_0^{R_M} r f(r)\, dr}. \tag{3.22}$$

Since $\mathcal{K}$ is a nonlinear operator, the differential equation for the AR estimate $\hat{f}$ of f_o is given by (3.20). There is a demonstration of the nonparametric AR corpuscle problem estimator in chapter 4, where we also show that the problem is essentially linear.

3.3 Poisson Process Intensity Estimation

In this section, we discuss the application of AR estimation to the Poisson process. Due to the close relationship between density estimation and Poisson process intensity estimation, previous results apply here with only slight modification.

With h fixed, the AR functional for estimation of the intensity $g_o = G_o'$ of a Poisson counting process N on $[0,1]$ is

$$J_h(g) = -\int_0^1 \frac{g}{h}\, dN + \frac{1}{2}\int_0^1 \frac{g^2}{h}.$$

The natural constrained optimization problem is then

$$\underset{g}{\text{minimize}}\ J_h(g)\ \text{ subject to }\ g \in L_2 \cap \mathcal{C}, \tag{3.23}$$

where the constraint set is $\mathcal{C} = \{g : g \geqslant 0\}$. We can construct a sequence $\{g_m\}_{m=1}^{\infty}$ in $L_2 \cap \mathcal{C}$ with the property that $J(g_m) \to -\infty$ as $m \to \infty$, thereby showing that J is unbounded. As a result, the optimization problem has no solution in L_2. In fact, $G_m(t) = \int_0^t g_m(u)\, du$ converges in L_2 to N, the sample path of the counting process. Details are in the proof of:

Theorem 3.12. *Problem* (3.23) *has no solution.*

As in the case of density estimation, we can penalize the objective to obtain a related problem that does have a useful solution. Specifically, the penalized problem is

$$\underset{g}{\text{minimize}}\ J_h(g)\ \text{ subject to }\ g \in \mathcal{H}_p \cap \mathcal{C}, \tag{3.24}$$

where

$$J_h(g) = -\int \frac{g}{h}\, dN + \frac{1}{2}\int \frac{g^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}g)^2}{h}$$

for some $\alpha > 0$. As in the case of density estimation, $\mathcal{D}$ is a linear differential operator of order $p \geqslant 1$. Of course, $\mathcal{D}$ has no constant term and is defined on a suitable domain. We formally state:

Theorem 3.13. *Problem* (3.24) *has a unique solution.*

If the differential operator $\mathcal{D}$ has order p, then the problem has a unique solution with $p-1$ continuous derivatives. Thus, one may choose the penalization based on which function (derivative of G_o) one wishes to estimate.

As in the parametric case, we have a theorem characterizing the limit of the estimator sequence.

Theorem 3.14. *With n fixed, if the AR estimator sequence associated with problem* (3.24) *converges, its limit is a maximum penalized likelihood estimator.*

3.3.1 Representation of the Intensity Estimator

In this section, we give a characterization of the AR intensity estimator. The situation is analogous to the density estimation case presented in section 3.1.1, so the discussion is brief.

The intensity estimator $\hat{g}$ is the solution of

$$J'(g)(r) = 0 \quad \text{for} \quad r \in \mathcal{H}_p,$$

and so the estimator satisfies the differential equation

$$\left(\mathcal{I} + \alpha h \cdot \mathcal{D}^* \frac{1}{h} \mathcal{D}\right) g = dN.$$

Here, $dN = \sum_{i=1}^{n} \delta_{t_i}$, where δ_t is the point mass at t. We can write

$$\hat{g}(t) = \sum_{i=1}^{n} s_{t_i}(t) \quad \text{and} \quad \hat{G}(t) = \int_0^t \hat{g}(u)\, du = \sum_{i=1}^{n} S_{t_i}(t),$$

where $\frac{d}{du} S_t(u) = s_t(u)$, and observe that s_t satisfies

$$\left(\mathcal{I} + \alpha h \cdot \mathcal{D}^* \frac{1}{h} \mathcal{D}\right) s_t = \delta_t.$$

3.3.2 Special Case

We can compute the solution of the AR intensity estimation problem in this special case, which is the Poisson process analogue of the problem of section 3.1.2.1.

With $\mathcal{D}x = x'$ and $h \equiv 1$, we can derive a closed-form solution for the unconstrained version of problem (3.24). We impose the boundary conditions $s_t'(0) = s_t'(1) = 0$ to specify a solution. We see that the solution does in fact satisfy the constraints. Since $H(t) = t$, $h(t) = 1$, and $h'(t) = 0$, we have

$$s_t - \alpha s_t'' = \delta_t.$$

Integration yields

$$S_t(u) - \alpha S_t''(u) = \begin{cases} 0, & u < t \\ 1, & u \geqslant t \end{cases}$$

with boundary conditions $S_t(0) = S_t''(0) = S_t''(1) = 0$, for $S_t \in \mathcal{H}_2$. As in section 3.1.2.1, the solution is

$$S_t(u) = \begin{cases} \operatorname{csch}\beta \cosh[\beta(1-t)] \sinh(\beta u), & u < t \\ 1 - \operatorname{csch}\beta \cosh(\beta t) \sinh[\beta(1-u)], & u \geqslant t, \end{cases}$$

where $\beta = \alpha^{-1/2}$. Also, note that $S_t(1) = 1$ for all $t \in [0,1]$. The nonparametric compensator estimate is then

$$\hat{G}(u) = \sum_{i=1}^{n} S_{t_i}(u).$$

This establishes existence and uniqueness of the solution, and the form of the unconstrained solution. Since

$$s_t(u) = S_t'(u) = \begin{cases} \beta \operatorname{csch}\beta \cosh[\beta(1-t)] \cosh(\beta u), & u < t \\ \beta \operatorname{csch}\beta \cosh(\beta t) \cosh[\beta(1-u)], & u \geqslant t, \end{cases}$$

we see that $s_t(u) \geqslant 0$ for all t and $u \in [0,1]$; therefore, the unconstrained intensity estimator $\hat{g} = \hat{G}'$ satisfies $g \geqslant 0$. So we have solved the constrained optimization problem. We recognize the solution $\hat{g}$ as a (boundary-corrected) bilateral exponential kernel intensity estimator.

3.4 Proofs

Proof of Theorem 3.1. Let $t_1, \ldots, t_n$ be ordered distinct data in $(0,1)$.

Suppose that $h(t_i) > 0$ for all i. This is reasonable because the "correct" value of h is f_o, which is necessarily positive at each t_i. Furthermore, we suppose that h is continuous since we are primarily interested in smooth (continuous or even differentiable) estimates of the density.

For m large enough so that t_1, $t_i - t_{i-1}$, and $1 - t_n$ are all greater than $1/m$, let

$$g_m(t) = \begin{cases} \frac{m}{n}, & (\exists k \in \{1,\ldots,n\})\ |t - t_k| \leqslant \frac{1}{2m} \\ 0, & (\forall k \in \{1,\ldots,n\})\ |t - t_k| > \frac{1}{2m}. \end{cases}$$

Clearly, $g_m \in L_2 \cap \mathcal{C}$ for all m large enough. Note that

$$J_{n,h}(g_m) = -\frac{m}{n^2} \sum_{i=1}^{n} \frac{1}{h(t_i)} + \frac{m^2}{2n^2} \sum_{i=1}^{n} \int_{t_i - \frac{1}{2m}}^{t_i + \frac{1}{2m}} \frac{1}{h(t)}\, dt.$$

Since

$$\int_{t_i - \frac{1}{2m}}^{t_i + \frac{1}{2m}} \frac{m}{h(t)}\, dt \to \frac{1}{h(t_i)} \quad \text{as} \quad m \to \infty,$$

it follows that

$$\frac{1}{m} J_{n,h}(g_m) \to -\frac{1}{2n^2} \sum_{i=1}^{n} \frac{1}{h(t_i)} \quad \text{as} \quad m \to \infty.$$

We have established that $J_{n,h}(g_m) \to -\infty$ as $m \to \infty$. Thus, J is unbounded below on $L_2 \cap \mathcal{C}$, and the optimization problem (3.1) has no solution. □

Proof of Lemma 3.2. The equivalence of $\|\cdot\|_{p,1}$ and $\|\cdot\|_{p,p}$ is a special case of Corollary 4.16 of Adams [1]. For the equivalence of $\|\cdot\|_{p,p}$ and $\|\cdot\|_{p,\mathcal{D}}$, see Silverman [66]. □

Proof of Theorem 3.3. See Thompson and Tapia [73]. □

Proof of Corollary 3.4. We verify the hypotheses of Theorem 3.3 in the context of problem (3.6), where the Hilbert space is $\mathcal{H} = \mathcal{H}_p$, the constraint set is

$$\mathcal{C} = \{f \in \mathcal{H}_p : f \geqslant 0 \text{ and } \int f = 1\},$$

and the functional is

$$J(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2}\int \frac{f^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}f)^2}{h}.$$

The constraint set $\mathcal{C}$ is clearly convex. To see that $\mathcal{C}$ is closed, consider a sequence g_n in $\mathcal{C}$ and a point $g \in \mathcal{H}$ with $\|g_n - g\| \to 0$. Since g and each of the g_n are absolutely continuous, we have $g_n \to g$ pointwise, and therefore $g \geqslant 0$. Writing $g = g_n - (g_n - g)$, we have $\int g = 1 - \int (g_n - g)$. Since $|\int (g_n - g)| \leqslant \int |g_n - g| \leqslant \sqrt{\int |g_n - g|^2} \leqslant \|g_n - g\|$, by Jensen's inequality, we have $\int g = 1$. Therefore, $\mathcal{C}$ is closed.

Again, since $\|f - g\| \to 0$ implies $f \to g$ pointwise, each map $f \mapsto f(t)/h(t)$ is continuous. The definition of the norm, along with the norm equivalences of Lemma 3.2, implies that $f \mapsto \int f^2/h$ and $f \mapsto \int (\mathcal{D}f)^2/h$ are continuous. Therefore, J is continuous.

The first Gâteaux derivative of $J(f)$ in the direction r is

$$J'(f)(r) = -\int \frac{r}{h}\, dF_n + \int \frac{fr}{h} + \alpha \int \frac{\mathcal{D}f\, \mathcal{D}r}{h}.$$

The second Gâteaux derivative of $J(f)$ in the directions r and s is

$$J''(f)(r,s) = \int \frac{rs}{h} + \alpha \int \frac{\mathcal{D}r\, \mathcal{D}s}{h}.$$

The cone tangent to $\mathcal{C}$ at f is defined as

$$T(f) = \{\eta \in \mathcal{H} : \exists t > 0, \text{ such that } f + t\eta \in \mathcal{C}\},$$

and, by definition, J'' is uniformly positive definite in $\mathcal{C}$ if there is a $k > 0$ such that for each $f \in \mathcal{C}$

$$J''(f)(\eta, \eta) \geqslant k\|\eta\|^2, \quad \forall \eta \in T(f).$$

In fact, by the norm equivalences of Lemma 3.2, we see that J'' is uniformly positive definite on all of $\mathcal{H}$. □

Proof of Theorem 3.5. The negative penalized loglikelihood for density estimation is

$$L(f) = -\int \log f \, dF_n + \alpha\nu(f)$$

for a suitable penalty functional ν. Its Gâteaux derivative is

$$L'(f)(r) = -\int \frac{r}{f}\, dF_n + \alpha\nu'(f)(r).$$

We express the constraint $\int f = 1$ as the functional relationship $Tf = 0$ and note that the derivative of T is

$$T'(f)(r) = \int r.$$

By the method of Lagrange multipliers for infinite-dimensional constrained optimization problems, the penalized likelihood estimator f satisfies for all r and some real λ

$$L'(f)(r) + \lambda T'(f)(r) = 0,$$

which is to say

$$-\int \frac{r}{f}\, dF_n + \lambda \int r + \alpha\nu'(f)(r) = 0.$$

See Luenberger [46] for a discussion of the Lagrange multiplier method in infinite-dimensional spaces. Now consider minimization of the AR objective functional

$$J(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2}\int \frac{f^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}f)^2}{h}.$$

Differentiating, we have

$$J'(f)(r) = -\int \frac{r}{h}\, dF_n + \int \frac{fr}{h} + \alpha \int \frac{\mathcal{D}f\, \mathcal{D}r}{h},$$

and the constrained solution f satisfies for all r and some real μ

$$J'(f)(r) + \mu T'(f)(r) = 0,$$

which is

$$-\int \frac{r}{h}\,dF_n + \int \frac{fr}{h} + \mu \int r + \alpha \int \frac{\mathcal{D}f\,\mathcal{D}r}{h} = 0.$$

Upon convergence, we have $f = h$, so the relationship is

$$-\int \frac{r}{f}\,dF_n + (1+\mu)\int r + \alpha \int \frac{\mathcal{D}f\,\mathcal{D}r}{f} = 0.$$

Identifying $\lambda = 1 + \mu$ and $\nu'(f)(r) = \int \frac{\mathcal{D}f\,\mathcal{D}r}{f}$ completes the proof. □

Proof of Theorem 3.6. See Bosq and Lecoutre [6]. □

Proof of Corollary 3.7. See Bosq and Lecoutre [6]. □

Proof of Lemma 3.9. See the proof of Theorem 3.11. □

Proof of Lemma 3.10. See the proof of Theorem 3.11. □

Proof of Theorem 3.11. The relevant norms expressed in terms of eigenfunction expansions are

$$\|f_o\|_w^2 = \sum_{i=1}^{\infty} c_i^2, \quad \|\Omega_{1,w} f_o\|_w^2 = \sum_{i=1}^{\infty} \lambda_i^{-2} c_i^2, \quad \|\mathcal{R}_{\alpha,w} f_o\|_w^2 = \sum_{i=1}^{\infty} \gamma_i^2 c_i^2,$$

$$\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|_w^2 = \sum_{i=1}^{\infty} \gamma_i^2 (\beta_i - c_i)^2, \quad \text{and} \quad \|f_o - \mathcal{R}_{\alpha,w} f_o\|_w^2 = \sum_{i=1}^{\infty} (1-\gamma_i)^2 c_i^2.$$

Then the error of the estimator can be bounded by

$$\begin{aligned} \tfrac{1}{2}\|\hat{f} - f_o\|_w^2 &\leqslant \|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|_w^2 + \|\mathcal{R}_{\alpha,w} f_o - f_o\|_w^2 \\ &= \sum_{i=1}^{\infty} \gamma_i^2 (\beta_i - c_i)^2 + \sum_{i=1}^{\infty} (1-\gamma_i)^2 c_i^2. \end{aligned}$$

To compute the expected value, note that the expectation of the first sum is a (weighted) sum of the $\operatorname{Var}\beta_i$. Let $\sigma_{ij}^2 = \int \psi_i \psi_j w^2\,dF_o$. It is straightforward to compute $\operatorname{Cov}(\beta_i, \beta_j) = n^{-1}(\sigma_{ij}^2 - c_i c_j)$, so that $\operatorname{Var}\beta_i = n^{-1}(\sigma_{ii}^2 - c_i^2)$. We then have the expectation

$$\mathrm{E}\,\|\hat{f} - \mathcal{R}_{\alpha,w} f_o\|_w^2 = \sum_{i=1}^{\infty} \gamma_i^2 \operatorname{Var}\beta_i = \frac{1}{n}\sum_{i=1}^{\infty} \gamma_i^2 (\sigma_{ii}^2 - c_i^2),$$

which gives a bound on the expected error.

$$\begin{aligned}\tfrac{1}{2}\,\mathrm{E}\,\|\hat{f}-f_o\|_w^2 &\leqslant \mathrm{E}\,\|\hat{f}-\mathcal{R}_{\alpha,w}f_o\|_w^2+\|\mathcal{R}_{\alpha,w}f_o-f_o\|_w^2\\ &=\frac{1}{n}\sum_{i=1}^{\infty}\gamma_i^2(\sigma_{ii}^2-c_i^2)+\sum_{i=1}^{\infty}(1-\gamma_i)^2c_i^2=n^{-1}S_1+S_2. \quad (3.25)\end{aligned}$$

In the analysis of S_1, we confine our attention to the case of $h = f_o$, or $w = 1/f_o$. Then we have $\sigma_{ij}^2 = \int \psi_i\psi_j w = \delta_{ij}$ and $\operatorname{Var}\beta_i = n^{-1}(1-c_i^2) \leqslant n^{-1}$, so that

$$S_1 \leqslant \sum_{i=1}^{\infty}\gamma_i^2 = \sum_{i=1}^{\infty}\frac{1}{(1-\alpha+\alpha\lambda_i^{-1})^2}.$$

We approximate S_1 with an integral in the manner of Silverman [66] or Wahba [78].

$$S_1 \sim \frac{1}{(1-\alpha)^2}\int_0^{\infty}\frac{1}{(1+\theta x^{2p})^2}\,dx = \frac{I_1}{(1-\alpha)^2},$$

where $\theta = \alpha/(1-\alpha)$. We use the change of variables

$$z=\frac{1}{1+\theta x^{2p}},\quad x=\left(\frac{1-z}{\theta z}\right)^{1/2p},\quad \text{and}\quad dx=-\frac{z^{1-1/2p}(1-z)^{1/2p-1}}{2p\,\theta^{1/2p}\,z^2}\,dz$$

to obtain

$$I_1=\frac{B\left(2-\frac{1}{2p},\frac{1}{2p}\right)}{2p\theta^{1/2p}}=O(\alpha^{-1/2p}).$$

Therefore,

$$n^{-1}S_1 = O\left(n^{-1}\alpha^{-1/2p}\right).$$

We can obtain a bound for S_2 by making use of the smoothness condition $\|\Omega_{1,w}f\|_w^2 = \sum_{i=1}^{\infty}\lambda_i^{-2}c_i^2 < \infty$. Observe that

$$\begin{aligned}S_2=\sum_{i=1}^{\infty}(1-\gamma_i)^2c_i^2&=\sum_{i=1}^{\infty}\frac{\alpha^2(\lambda_i-1)^{-2}c_i^2}{(1-\alpha+\alpha\lambda_i^{-1})^2}\\ &\leqslant\alpha^2\sum_{i=1}^{\infty}\frac{\lambda_i^{-2}c_i^2}{(1-\alpha+\alpha\lambda_i^{-1})^2}=\theta^2\sum_{i=1}^{\infty}\frac{\lambda_i^{-2}c_i^2}{(1+\theta\lambda_i^{-1})^2},\end{aligned}$$

and that

$$\frac{\lambda_i^{-2} {c_i}^2}{(1+\theta\lambda_i^{-1})^2} \nearrow \lambda_i^{-2} {c_i}^2 \quad \text{as} \quad \theta \searrow 0, \quad \forall i.$$

Hence, by the dominated convergence theorem (the dominator is also the limit here),

$$S_2 \to \theta^2 \|\Omega_{1,w} f\|_w^2 \text{ as } \theta \to 0.$$

Therefore

$$S_2 = O\left(\alpha^2\right).$$

We thus obtain the asymptotic rate of convergence for expected error

$$\mathrm{E}\,\|\hat{f} - f_o\|_w^2 = O\left(n^{-1}\alpha^{-1/2p} + \alpha^2\right).$$

In the native Sobolev norm given by

$$\|x\|_{1,w}^2 = \|x\|_w^2 + \|\mathcal{D}x\|_w^2 = \langle x, \Omega_{1,w} x\rangle_w,$$

we have

$$\|x\|_{1,w}^2 = \sum_{i=1}^{\infty} \lambda_i^{-1} \langle x, \psi_i\rangle_w^2.$$

As in equation (3.25), we can write

$$\tfrac{1}{2}\,\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 \leqslant \frac{1}{n}\sum_{i=1}^{\infty} \lambda_i^{-1}\gamma_i^2(\sigma_{ii}^2 - c_i^2) + \sum_{i=1}^{\infty} \lambda_i^{-1}(1-\gamma_i)^2 c_i^2 = \tfrac{1}{n} S_1 + S_2.$$

And when $w = 1/f$, we have

$$S_1 \leqslant \sum_{i=1}^{\infty} \lambda_i^{-1}\gamma_i^2 = \sum_{i=1}^{\infty} \frac{\lambda_i^{-1}}{(1-\alpha+\alpha\lambda_i^{-1})^2}.$$

Approximating S_1 with an integral results in

$$S_1 \sim \frac{1}{(1-\alpha)^2}\int_0^{\infty} \frac{x^{2p}}{(1+\theta x^{2p})^2}\,dx = \frac{1}{(1-\alpha)^2} I_1.$$

We evaluate

$$I_1 = \frac{B\left(1-\frac{1}{2p}, 1+\frac{1}{2p}\right)}{2p\,\theta^{1+1/2p}}$$

to obtain

$$S_1 = O\left(\alpha^{-1-1/2p}\right).$$

The second sum satisfies

$$S_2 \leqslant \theta^2 \sum_{i=1}^{\infty} \frac{\lambda_i^{-3} {c_i}^2}{(1+\theta\lambda_i^{-1})^2}.$$

Using $\|\Omega_{1,w} f_o\|_w^2 < \infty$, it can be shown by dominated convergence that $S_2 = O(\alpha)$. See Silverman [66]. Consequently,

$$\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 = O\left(n^{-1}\alpha^{-1-1/2p} + \alpha\right).$$

If we assume the additional smoothness condition $\|\Omega_{1,w} f_o\|_{1,w}^2 < \infty$ on the true density, we can apply dominated convergence to obtain $S_2 \to \theta^2 \|\Omega_{1,w} f\|_{1,w}^2 = O\left(\alpha^2\right)$ and in turn

$$\mathrm{E}\,\|\hat{f} - f_o\|_{1,w}^2 = O\left(n^{-1}\alpha^{-1-1/2p} + \alpha^2\right).$$

□

Proof of Theorem 3.12. Apply the method in the proof of Theorem 3.1.

□

Proof of Theorem 3.13. Apply the method in the proof of Corollary 3.4.

□

Proof of Theorem 3.14. Similar to Theorem 3.5. The negative penalized loglikelihood for a Poisson process on $[0,1]$ is

$$L(g) = -\int_0^1 \log g \, dN + \int_0^1 g + \alpha\nu(g)$$

for a suitable penalty functional ν. To evaluate the Gâteaux derivative, note that

$$\frac{d}{d\lambda} \int_0^1 \log(g + \lambda r)\, dN \bigg|_{\lambda=0} = \int_0^1 \frac{r}{g}\, dN$$

and

$$\frac{d}{d\lambda}\int_0^1 (g+\lambda r)\Big|_{\lambda=0} = \int_0^1 r.$$

Therefore, a penalized loglikelihood estimate g satisfies

$$L'(g)(r) = -\int_0^1 \frac{r}{g}\,dN + \int_0^1 r + \alpha\nu'(g)(r) = 0, \quad \forall r.$$

Now consider minimization of the objective

$$J(g) = -\int \frac{g}{h}\,dN + \frac{1}{2}\int \frac{g^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}g)^2}{h}.$$

Differentiating, we have

$$J'(g)(r) = -\int \frac{r}{h}\,dN + \int \frac{gr}{h} + \alpha\int \frac{\mathcal{D}g\,\mathcal{D}r}{h} = 0, \quad \forall r.$$

At a fixed point, we have $h = g$, so that

$$-\int \frac{r}{g}\,dN + \int r + \alpha\int \frac{\mathcal{D}g\,\mathcal{D}r}{g} = 0, \quad \forall r.$$

Thus, the methods produce the same solution if $\nu'(g)(r) = \int \frac{\mathcal{D}g\,\mathcal{D}r}{g}$. □

4. Practical Nonparametric AR Estimation

In this chapter, we consider the practical computational aspects of nonparametric asymptotic regression estimation. We have seen that there are closed-form solutions in certain cases for the nonparametric AR estimation problem. In general, however, this is not possible. So, in order to obtain numerical results, we use a finite-dimensional (discretized) approximation to the infinite-dimensional problem.

In section 4.1, we present the discretization scheme for density estimation and inverse problems. Since density estimation techniques apply to the problem of intensity estimation for completely observed Poisson processes, the Poisson process problem is not discussed explicitly. Section 4.2 describes a data-driven procedure for selection of the smoothing parameter. In section 4.3, by means of a Monte-Carlo simulation study, we compare nonparametric AR estimation to several competitive methods for solving the deconvolution problem.

4.1 Discretization Techniques

4.1.1 Density Estimation

We begin by recalling the nonparametric (penalized) AR density estimation problem from chapter 3. That is,

$$\underset{f}{\text{minimize}}\ J(f)\ \text{ subject to }\ f \in \mathcal{H}_p \cap \{f : f \geqslant 0, \textstyle\int f = 1\},$$

where the objective functional is

$$J(f) = -\int \frac{f}{h}\, dF_n + \frac{1}{2}\int \frac{f^2}{h} + \frac{\alpha}{2}\int \frac{(\mathcal{D}f)^2}{h}, \tag{4.1}$$

$\alpha > 0$ is a constant called the smoothing parameter, h is a p.d.f., $\mathcal{D}$ is a linear differential operator of order $p \geqslant 1$ with no constant term, and F_n is the empirical c.d.f. based on a sample $\mathbf{X} = \{X_1, \ldots, X_n\}$ of size n

distributed with true density f_o. In terms of the weighted L_2 inner product $\langle x, y \rangle_w = \int wxy$, we use $w = 1/h$ to express (4.1) as

$$\begin{aligned} J(f) &= -\langle f, f_n \rangle_w + \tfrac{1}{2}\|f\|_w^2 + \tfrac{\alpha}{2}\|\mathcal{D}f\|_w^2 \\ &= \tfrac{1}{2}\left\langle f, (\mathcal{I} + \alpha \tfrac{1}{w}\mathcal{D}^* w \mathcal{D}) f \right\rangle_w - \langle f_n, f \rangle_w \\ &= \tfrac{1}{2}\langle f, (w + \alpha \mathcal{D}^* w \mathcal{D}) f \rangle - \langle w f_n, f \rangle \\ &= \tfrac{1}{2}\langle f, \mathcal{Q} f \rangle - \langle b_n, f \rangle , \end{aligned}$$

where $\mathcal{Q} = w + \alpha \mathcal{D}^* w \mathcal{D}$ and $b_n = w f_n$. The discrete version of this problem is

$$\underset{f}{\text{minimize}}\ \tfrac{1}{2} f^t Q f - b_n^t f \ \text{ subject to } \ f \geqslant 0 \ \text{ and } \ k^t f = 1, \qquad (4.2)$$

which, of course, is the quadratic programming problem. High-quality software for solving this problem is readily available.

Note that the solution of the unconstrained problem

$$\underset{f}{\text{minimize}}\ \tfrac{1}{2} f^t Q f - b_n^t f$$

is also the solution of the system of linear equations

$$Qf = b_n. \qquad (4.3)$$

In some applications, the constraints can be ignored and the estimator can be computed as the solution of (4.3). An easier computational problem does not exist.

At an intermediate level of complexity, we have the equality-constrained problem

$$\underset{f}{\text{minimize}}\ \tfrac{1}{2} f^t Q f - b_n^t f \ \text{ subject to } \ k^t f = 1. \qquad (4.4)$$

The solution to (4.4) integrates to 1, but may attain negative values. However, in a reasonable proportion of practical situations, the solution does indeed satisfy the positivity constraint. If necessary, the solution can be

truncated and renormalized to obtain a non-negative density estimate. This observation, coupled with the complexity of (4.2) relative to (4.4), makes it worthwhile to compute (4.4), check for positivity, and solve the quadratic program only when necessary. In fact, (4.4) can be realized by solving two linear systems, as we now show.

Let $b = b_n$. The Lagrange multiplier condition for (4.4) is

$$Qf - (b + \lambda k) = 0$$

where λ is a real constant. With c and d such that $Qc = b$ and $Qd = k$, we have $f = c + \lambda d$. Because $k'f = k'(c + \lambda d) = 1$, we have $\lambda = (1 - k'c)/(k'd)$ and

$$f = c + \frac{1 - k'c}{k'd}\, d.$$

We now discuss the particular form of the (approximate) discrete representation. Our discretization is based on m values. Let D denote the $m \times m$ matrix representer of the differential operator $\mathcal{D}$, and let W denote the $m \times m$ diagonal weight matrix of the inner product. Here, f_n is a discrete representation of the empirical point measure (i.e., a histogram estimate with m bins), and $b_n = Wf_n$. The quantities g and k are m-vectors, and $Q = W + \alpha D^* W D$.

For convenience, we take the domain of all functions to be $[0, 1]$. Partition this interval into m subintervals of length $1/m$. Denote by I_i the i^{th} subinterval, so

$$I_1 = \left[\, 0\,,\, \frac{1}{m} \,\right] \quad \text{and} \quad I_i = \left(\, \frac{i-1}{m}\,,\, \frac{i}{m} \,\right], \; i \in \{2, \ldots, m\}.$$

Functions are constant on subintervals, with values to be taken at the midpoints of subintervals. The discrete domain values are then

$$(x_1, \ldots, x_m) \;\text{ where }\; x_i = \frac{i - 1/2}{m}, \; i \in \{1, \ldots, m\}.$$

The functions h, $w = 1/h$, and f are represented by m-vectors. I.e.,

$$h = (h_1, \ldots, h_m), \;\text{ where }\; h_i = h(x_i) \;\text{ for }\; i \in \{1, \ldots, m\},$$

and likewise for w and f. The vector k is the representer of $z \to \int z$, so

$$k = (k_1, \ldots, k_m), \quad \text{where} \quad k_i = \frac{1}{m} \quad \text{for} \quad i \in \{1, \ldots, m\}.$$

The discrete version of f_n is a histogram density estimator. Specifically, we have

$$f_n = (f_{n,1}, \ldots, f_{n,m}), \quad \text{where} \quad f_{n,i} = \frac{m}{n}|\mathbf{X} \cap I_i| \quad \text{for} \quad i \in \{1, \ldots, m\}. \quad (4.5)$$

The diagonal weight matrix W has entries $W_{ii} = w_i$. Derivative operators are represented by difference matrices. Let D_p denote the representer of $z \to z^{(p)}$. The first difference is

$$D_1 = m \begin{bmatrix} 1 & 0 & & & & \\ -1 & 1 & 0 & & & \\ & -1 & 1 & & & \\ & & -1 & \ddots & 0 & \\ & & & & 1 & 0 \\ & & & & -1 & 1 \end{bmatrix},$$

and higher-order differences are given by

$$D_p = \begin{cases} D_1^t D_{p-1}, & p \text{ even} \\ D_{p-1} D_1, & p \text{ odd}, \end{cases}$$

with the appropriate corrections for boundary conditions. These representations are used to compute solutions for any of the problems (4.2), (4.3), and (4.4).

When we refer to a smoothing parameter value of α, we are actually using α^p, where p is the order of the penalty operator. This makes the smoothing parameter independent of the penalty order.

The recursive sequence of AR estimators $(\hat{f}_0, \hat{f}_1, \hat{f}_2, \ldots)$ is computed by using an initial value of $\hat{f}_0$, and then setting $w = 1/\hat{f}_i$ to obtain $\hat{f}_{i+1}$ as the solution of (4.2), (4.3), or (4.4), as appropriate.

The following example graphs (Figures 4.1–4.5) illustrate the various solution options for AR estimation. We used S-PLUS to perform most of the numerical calculations. An IMSL routine was used to solve quadratic programs. One data set, the "Buffalo snowfall" data, was used in all the examples. This data set of size $n = 63$ consists of annual snowfall amounts in Buffalo, New York, for the winters of 1910/11 through 1972/73. This is a classic example of a data set that may come from a trimodal distribution.

Figure 4.1 shows the effect of variation in m, the discretization grid size. Values used were $m = 10$, 20, 50, 100, 250, and 500. These are single-step AR estimates with the penalty operator $\mathcal{D}z = z''$, a smoothing parameter value of $\alpha = 0.001$, and a uniform reference distribution. Solutions were computed using problem (4.4). The solutions are non-negative everywhere, so they are also solutions of problem (4.2). The graphs in Figure 4.1 explicitly show that the approximate solutions are piecewise constant. In the remaining figures, estimate values at the midpoints of intervals are connected with straight lines to give continuous graphs.

Figure 4.2 shows the effect of variation in α, the smoothing parameter. Values used were $\alpha = 0.0002$, 0.0005, 0.001, 0.002, 0.005, and 0.001. These are single-step AR estimates with the penalty operator $\mathcal{D}z = z''$, a grid size of $m = 100$, and a uniform reference distribution. Solutions were computed using problem (4.4).

Figure 4.3 shows the effect of various penalty operator orders. Here we used $\mathcal{D}z = z^{(p)}$ for $p = 1$, 2, 3, 4, 5, and 6. These are single-step AR estimates with a uniform reference distribution, a grid size of $m = 100$, and a smoothing parameter value of $\alpha = 0.001$. Solutions were computed using problem (4.2).

Figures 4.4 and 4.5 display recursive ARE sequences with five iterations. In all cases, we used a uniform initial weight, a grid size of $m = 100$, and a smoothing parameter value of $\alpha = 0.001$. In Figure 4.4, the penalty operator is $\mathcal{D}z = z''$, and Figure 4.5 depicts $\mathcal{D}z = z^{(p)}$ with $p = 1$, 2, 3, 4, 5, and 6. Solutions were computed using problem (4.2).

4.1.2 Inverse Problems

Recall the problem formulation from section 3.2. The random variables $X_1, \ldots, X_n$ are i.i.d. with unknown p.d.f. f_o, which we wish to estimate. Observed data are $Z_1, \ldots, Z_n$, where the p.d.f. of the Z_i is

$$g_o(t) = [\mathcal{K} f_o](t)$$

for some known operator $\mathcal{K}$. Here, we assume that $\mathcal{K}$ is a linear operator. The empirical c.d.f. based on the observable data is G_n, and the corresponding empirical point measure is g_n. An estimate f of f_o is a solution of the problem

$$\underset{f}{\text{minimize}}\ J(f) \ \text{ subject to } \ f \in \mathcal{H}_p \cap \left\{ f : f \geqslant 0, \textstyle\int f = 1 \right\}. \tag{4.6}$$

The objective functional is

$$\begin{aligned} J(f) &= -\langle g_n, \mathcal{K} f \rangle_w + \tfrac{1}{2} \|\mathcal{K} f\|_w^2 + \tfrac{\alpha}{2} \|\mathcal{D} f\|_w^2 \\ &= \tfrac{1}{2} \langle f, (\mathcal{K}^* w \mathcal{K} + \alpha \mathcal{D}^* w \mathcal{D}) f \rangle - \langle \mathcal{K}^* w g_n, f \rangle \\ &= \tfrac{1}{2} \langle f, \Omega f \rangle - \langle b_n, f \rangle . \end{aligned}$$

Let K denote the $m \times m$ matrix representer of the operator $\mathcal{K}$. The discrete versions of problem (4.6) are given by (4.2), (4.3), and (4.4) with $\Omega = \mathcal{K}^* w \mathcal{K} + \alpha \mathcal{D}^* w \mathcal{D}$ and $b_n = \mathcal{K}^* w f_n$. The only difference is the presence of the operator $\mathcal{K}$, so formulation of a linear inverse problem requires the additional determination of K. We now do this for two particular problems.

4.1.2.1 The Deconvolution Problem. Recall from section 3.2.1 that the convolution operator $\mathcal{K}$ is defined by

$$g(t) = [\mathcal{K} f](t) = \int k(t - x) f(x)\, dx,$$

where k is a known p.d.f. The discrete version is

$$g_j = \sum_{i=1}^{m} k_{j-i} f_i,$$

which is, in terms of its components,

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_m \end{bmatrix} = \begin{bmatrix} k_0 & k_{-1} & k_{-2} & \cdots & k_{-m+1} \\ k_1 & k_0 & k_{-1} & & k_{-m+2} \\ k_2 & k_1 & k_0 & & k_{-m+3} \\ \vdots & & & \ddots & \vdots \\ k_{m-1} & k_{m-2} & k_{m-3} & \cdots & k_0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_m \end{bmatrix}.$$

In the discrete representation with m intervals on $[0,1]$, we take f constant on subintervals with boundary points i/m, where $i \in \{0,\ldots,m\}$. To obtain the correct proportions in the representer K, we take k to be constant on intervals of width $1/m$ that are *centered* at the points i/m, $i \in \mathbb{Z}$. This leads to the definition

$$k_i = \int_{(i-1/2)/m}^{(i+1/2)/m} k(t)\,dt.$$

4.1.2.2 The Corpuscle Problem. Recall the formulation of the corpuscle problem from section 3.2.2. Spheres with random radii are distributed at random uniformly in a solid medium. The sphere radius p.d.f. is f_o, with support $[0, R_M]$. A slice through the medium gives data that are circles (sphere-slice intersections) with p.d.f. g_o. The functional $\mathcal{K}_o$, which describes the relationship between f_o and g_o, is nonlinear. Specifically,

$$g_o(t) = [\mathcal{K}_o f_o](t) = \frac{t\int_t^{R_M}(x^2-t^2)^{-1/2} f_o(x)\,dx}{\int_0^{R_M} x f_o(x)\,dx}.$$

However, we can define a new function f_* by $f_*(t) = f_o(t) \Big/ \int_0^{R_M} x f_o(x)\,dx$, and observe that

$$g_o(t) = [\mathcal{K}_o f_o](t) = [\mathcal{K} f_*](t) = t\int_t^{R_M}(x^2-t^2)^{-1/2} f_*(x)\,dx,$$

where $\mathcal{K}$ is a linear operator. Since f_* is simply a constant multiple of the p.d.f. f_o, which we wish to estimate, we can recover f_o by normalizing f_*;

i.e.,

$$f_o(t) = \frac{f_*(t)}{\int_0^{R_M} f_*(x)\,dx}.$$

To compute the discrete operator, let us first scale $[0, R_M]$ into $[0,1]$ by using $t = R_M s$ and $x = R_M z$. The integral then becomes

$$g(R_M s) = R_M s \int_s^1 (z^2 - s^2)^{-1/2} f(R_M z)\,dz.$$

The domain of integration proceeds from the midpoint of an interval to 1; thus, the discrete version is

$$g_i = R_M \cdot \frac{i - 1/2}{m} \left[f_i \int_{\frac{i-1/2}{m}}^{\frac{i}{m}} \frac{dz}{\sqrt{z^2 - \left(\frac{i-1/2}{m}\right)^2}} + \sum_{j=i+1}^{m} f_j \int_{\frac{j-1}{m}}^{\frac{j}{m}} \frac{dz}{\sqrt{z^2 - \left(\frac{i-1}{m}\right)^2}} \right].$$

Since $\int (z^2 - s^2)^{-1/2}\,dz = \log\left[z + (z^2 - s^2)^{1/2}\right]$, we have

$$g = Kf, \quad \text{or} \quad g_i = R_M \cdot \tfrac{i-1/2}{m} \sum_{j=1}^{m} v_{ij} f_j \quad \text{for} \quad i \in \{1, \ldots, m\},$$

where

$$v_{ij} = \begin{cases} 0, & j < i \\ \log\left[\frac{i + \sqrt{i^2 - (i-1/2)^2}}{i-1/2}\right], & j = i \\ \log\left[\frac{j + \sqrt{j^2 - (i-1/2)^2}}{j-1 + \sqrt{(j-1)^2 - (i-1/2)^2}}\right], & j > i. \end{cases}$$

4.2 Selecting the Smoothing Parameter

Practically speaking, we need an automatic procedure for selecting the smoothing parameter. Cross-validation is suited to least-squares problems and has been applied to spline smoothing and other statistical estimation and regression problems. Since discrete AR estimation is a smoothing spline problem, the technique of cross-validation is directly applicable. We use

generalized cross-validation (GCV) to select the AR smoothing parameter in density estimation and linear inverse problems. For background information on cross-validation and GCV, see Wahba [78] and [79], Gu [23], and Härdle [27].

4.2.1 Density Estimation

The discrete representation (4.3) has unconstrained solution $g = M_\alpha f_n$, where $M_\alpha = (W + \alpha D^* W D)^{-1} W$. Let Tr denote the trace operator, and let the weighted discrete inner product $\langle \cdot, \cdot \rangle_W$ be given by $\langle a, b \rangle_W = \sum_{i=1}^m a_i b_i w_i$.

The GCV score function

$$C(\alpha) = \frac{\|(I - M_\alpha) f_n\|_W^2}{[\mathrm{Tr}(I - M_\alpha)]^2}$$

is an estimator of mean-squared error. The GCV criterion for selection of the smoothing parameter α is

$$\underset{\alpha}{\text{minimize}}\ C(\alpha).$$

Note that the numerator of $C(\alpha)$ is simply the weighted squared deviation $\sum_{i=1}^m (g_i - f_{n,i})^2 w_i$, where $(f_{n,1}, \ldots, f_{n,m})$ is the histogram density estimator of (4.5).

For an illustrative example, we take a sample of size $n = 100$ from the beta distribution (defined on page 97) with density $\beta(\cdot, 3, 5)$ and compute the GCV score for a range of α values (Figure 4.6). Using the value of α selected by the GCV criterion, we then compute the second iteration of the AR sequence (Figure 4.7).

4.2.2 Inverse Problems

This is similar to the standard density estimation case. The discrete representation has unconstrained solution $f = M_\alpha g_n$, where $M_\alpha = (K^* W K + \alpha D^* W D)^{-1} K^* W$. The technique of GCV can be adapted to the linear inverse

problem. The GCV criterion in this case is

$$\underset{\alpha}{\text{minimize}}\ C(\alpha) = \frac{\|(I - KM_\alpha)g_n\|_W^2}{[\text{Tr}(I - KM_\alpha)]^2}.$$

There is an extra K in the score because $M_\alpha g_n$ is an estimator of f and $KM_\alpha g_n$ is an estimator of g. Note that g is the distribution of the (observable) data.

The numerator of $C(\alpha)$ in this case is the weighted squared deviation $\sum_{i=1}^m [(Kf)_i - g_{n,i}]^2 w_i$, where Kf is the transformed estimator of the underlying unobservable density, hence an estimator of the observable data density g. Of course, $(g_{n,1}, \ldots, g_{n,m})$ is the histogram density estimator of the observable data.

For an example of GCV in deconvolution estimation, we take a sample of size $n = 100$ from the density $\beta(\cdot, 3, 5)$ as the signal. The noise distribution is normal with a standard deviation of 0.1 and zero mean. Once again, we compute the GCV score for a range of α values (Figure 4.8). Using the value of α selected by the GCV criterion, we then compute the second iteration of the AR sequence (Figure 4.9).

For an example of GCV in corpuscle problem estimation, we take a sample of size $n = 250$ from density $\beta(\cdot, 5, 3)$ as the signal. As usual, we compute the GCV score for a range of α values (Figure 4.10). Using the α selected by the GCV criterion, we then compute the second iteration of the AR sequence (Figure 4.11).

4.3 Simulation Study: Deconvolution

In this section, we present the results of a modest simulation study of the deconvolution problem, in which we compare the the AR deconvolution estimator to the NEMS estimator of Eggermont and LaRiccia [15] and the Fourier kernel method studied by Stefanski [69], Fan [18], and others.

We use the signal and noise distributions for which results are tabulated in Eggermont and LaRiccia [15] and compare our AR results to theirs for

the NEMS and Fourier estimators. Optimal smoothing-parameter selection is possible for all three estimators, since we know the true signal distributions. There is an automatic procedure for selecting the NEMS smoothing parameter, and we use GCV to select the AR smoothing parameter. All three estimators are compared for optimal selection of the smoothing parameter, and the NEMS and AR estimators are compared for automatic selection. The basis for comparing the various estimators $\hat{f}$ in this simulation is the L_1 error

$$\|\hat{f} - f\|_1 = \int |\hat{f} - f|,$$

where f is the density of the underlying signal (without noise). The average L_1 error for a number of repetitions is taken as an estimate of $\mathrm{E}\,\|\hat{f} - f\|_1$.

Average L_1 errors for various signal distributions are presented in Tables 4.1 (with normal noise) and 4.2 (with uniform noise) for a sample size of 100. Sample means and standard deviations of L_1 errors for a range of sample sizes are presented in Table 4.3, but these results are limited to two of the normal-noise cases.

For the simulation study, signal and noise distributions are based on the normal density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{and} \quad \phi(x;\, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x}{\sigma}\right),$$

the uniform density

$$u(x) = I_{[0,1]}(x) \quad \text{and} \quad u(x;\, \sigma) = \frac{1}{\sigma} u\left(\frac{x}{\sigma}\right),$$

and the beta density

$$\beta(x,\, a,\, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} I_{[0,1]}(x) \quad \text{and} \quad \beta(x,\, a,\, b;\, \sigma) = \frac{1}{\sigma}\beta\left(\frac{x}{\sigma},\, a,\, b\right).$$

The six signal densities we use in the simulation are

$$
\begin{aligned}
&(D1) \quad f(x) = 0.9\phi(x-5;\,0.5) + 0.1\phi(x-7;\,0.25),\\
&(D2) \quad f(x) = 0.9\phi(x-5;\,0.5) + 0.1\phi(x-7;\,0.5),\\
&(D3) \quad f(x) = 0.5\phi(x-5.82;\,0.57) + 0.5\phi(x-4.18;\,0.57),\\
&(D4) \quad f(x) = \phi(x-5),\\
&(D5) \quad f(x) = u(x-3;\,5), \quad \text{and}\\
&(D6) \quad f(x) = \beta(x-3,\,2.4,\,3.6;\,5).
\end{aligned}
$$

Convolution is accomplished, of course, by adding noise to the signal. Noise densities used in the simulation are normal (with various standard deviations) and uniform on $[0,1]$. Combinations of signal and normal noise are referenced as N1–N8. See Table 4.1 for the particular values. The corresponding combinations with uniform noise (referenced as U1, etc.) are detailed in Table 4.2.

The NEMS and Fourier estimation parameters are detailed in Eggermont and LaRiccia [15]. We now describe the AR parameters for this simulation.

For the given combinations of signal and noise, an estimation domain of $[0,10]$ is adequate to capture most of the data. Observations fall outside of this region only rarely.

The AR penalty $\mathcal{D}x = x''$ is used in all cases except N5, U5, and N7, in which $\mathcal{D}x = x'$ is used for reasons of numerical stability.

AR estimators are calculated using model (4.4) of page 88 with an iteration count of $i = 2$. The initial weight for the AR procedure is the constant function. In the case of automatic smoothing parameter selection, the result of GCV applied to the first-step AR estimator is used as the second-step smoothing parameter, in addition to using the first-step estimate as the reference measure in the second iteration.

Sample sizes are $n = 100$ in Tables 4.1 and 4.2, and n =50, 100, 250, 500, and 1000 in Table 4.3. A simulation repetition count of $N = 1000$ is used in all cases. To conserve computation time, a grid size of $m = 100$ is used throughout.

The S-PLUS language, version 3.4, is used for programming the procedure, along with a few utility functions written in the C language. A native S-PLUS minimization routine is used in both the optimal and GCV smoothing parameter selection. This routine searches for local extrema on an interval. Reasonable α-search regions are selected separately for each case. It is interesting to note that the GCV search is more efficient than the optimal α search.

The computing platform used is a Silicon Graphics SGI-4D/PCXL-8 with 24 200-MHz IP19 processors and 1 GB of main memory. In this environment, the simulation runs in about 24 hours.

4.3.1 Observations

For the study with normal noise and $n = 100$ (Table 4.1), the NEMS and AR estimators are competitive, whereas the Fourier estimator has much larger errors. With optimal smoothing-parameter selection, the NEMS error exceeds the AR error in six out of eight cases; and with automatic selection, the AR error exceeds the NEMS error in four out of eight cases. The Fourier estimator has much higher errors.

The NEMS estimator performs slightly better than the AR estimator in the case of uniform noise (Table 4.2). With optimal smoothing-parameter selection, the NEMS error exceeds the AR error in two out of six cases; and with automatic selection, the AR error exceeds the NEMS error in five out of six cases.

Several interesting observations arise from consideration of the results in Table 4.3. Since various sample sizes are tested here, the data provide an empirical indication of the rates of L_1 error convergence for the NEMS and AR estimators in the cases of optimal and automatic smoothing-parameter selection. This is limited, of course, to the two signal and noise combinations under study here.

Error variances are comparable, with the AR estimator having marginally lower dispersion for optimal selection, and the NEMS estimator having slightly lower dispersion for automatic selection.

See Figures 4.12 and 4.13 for graphical presentations of the mean error rates. Expected L_1 error can be modeled as

$$\mathrm{E}\,\|\hat{f} - f\|_1 = k\, n^{-p},$$

and the coefficients k and p can be obtained from the Table 4.3 results by linear regression. Estimates of the coefficients are presented in Table 4.4.

The Fourier estimator has higher error and a slower rate of convergence. With optimal selection, the AR estimator has lower errors than the NEMS estimator for both distributions tested. Also, the AR error converges at a faster rate.

With automatic selection, convergence rates of the AR and NEMS estimators are practically the same. The NEMS error is slightly lower for distribution N5, and the AR estimator has a slightly better rate for N2.

The AR estimator (for N2 and N5) and the NEMS estimator (for N2) have optimal selection error convergence rates that are faster than the corresponding automatic rates, as would be expected. The small differences in slope indicate that the automatic-selection procedures give estimators that almost achieve the optimal-selection rate. For these cases, note that the lines in Figure 4.13 are practically parallel.

In conclusion, asymptotic regression provides a reasonable and practical technique for deconvolution estimation.

Table 4.1. Mean L_1 Error for Fourier, NEMS, and AR Deconvolution Estimators with Normal $\phi(\cdot\,;\,\sigma)$ Noise, $n = 100$

	f	σ	optimal α			automatic α	
			Fourier	NEMS	AR	NEMS	AR
N1	(D2)	.50	.389	.210	.210	.326	.250
N2	(D2)	.29	.280	.167	.154	.205	.205
N3	(D3)	.58	.301	.201	.205	.254	.258
N4	(D4)	.50	.242	.131	.105	.157	.156
N5	(D4)	.29	.195	.127	.116	.144	.160
N6	(D5)	.29	.265	.231	.226	.246	.287
N7	(D6)	.29	.209	.136	.121	.150	.162
N8	(D1)	.29	.298	.181	.168	.225	.213

Table 4.2. Mean L_1 Error for NEMS and AR Deconvolution Estimators with Uniform $u(\cdot\,;\,1)$ Noise, $n = 100$

	f	optimal		automatic	
		NEMS	AR	NEMS	AR
U1	(D2)	.169	.169	.194	.228
U3	(D3)	.168	.170	.189	.224
U4	(D4)	.133	.111	.170	.162
U6	(D5)	.232	.236	.269	.289
U7	(D6)	.134	.147	.164	.182
U8	(D1)	.181	.177	.209	.231

Table 4.3. Mean and Standard Deviation of L_1 Error for Fourier, NEMS, and AR Deconvolution Estimators with Normal $\phi(\cdot\ ;\ .29)$ Noise, $n = 50$, 100, 250, 500, and 1000

		optimal α						automatic α			
		Fourier		NEMS		AR		NEMS		AR	
	n	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
N2	50	.312	.079	.217	.077	.202	.078	.254	.083	.268	.108
	100	.280	.060	.167	.060	.154	.058	.205	.067	.205	.090
	250	.244	.045	.120	.042	.108	.038	.147	.048	.144	.067
	500	.222	.033	.094	.032	.081	.029	.114	.037	.112	.053
	1000	.204	.025	.071	.023	.062	.021	.086	.027	.086	.038
N5	50	.227	.072	.163	.076	.154	.063	.196	.082	.208	.092
	100	.195	.051	.126	.053	.116	.047	.144	.053	.160	.066
	250	.163	.036	.093	.036	.079	.030	.103	.036	.112	.041
	500	.142	.026	.074	.026	.061	.021	.076	.026	.086	.028
	1000	.127	.020	.055	.018	.045	.014	.059	.019	.062	.017

Table 4.4. Empirical L_1 Error Rate Coefficients for Fourier, NEMS, and AR Deconvolution Estimators in the Model $\mathrm{E}\|\hat{f} - f\|_1 = k\, n^{-p}$

		optimal α			automatic α	
		Fourier	NEMS	AR	NEMS	AR
N2	k	0.541	0.920	0.955	1.069	1.176
	p	0.143	0.369	0.396	0.362	0.379
N5	k	0.481	0.655	0.768	0.925	1.005
	p	0.195	0.355	0.411	0.400	0.400

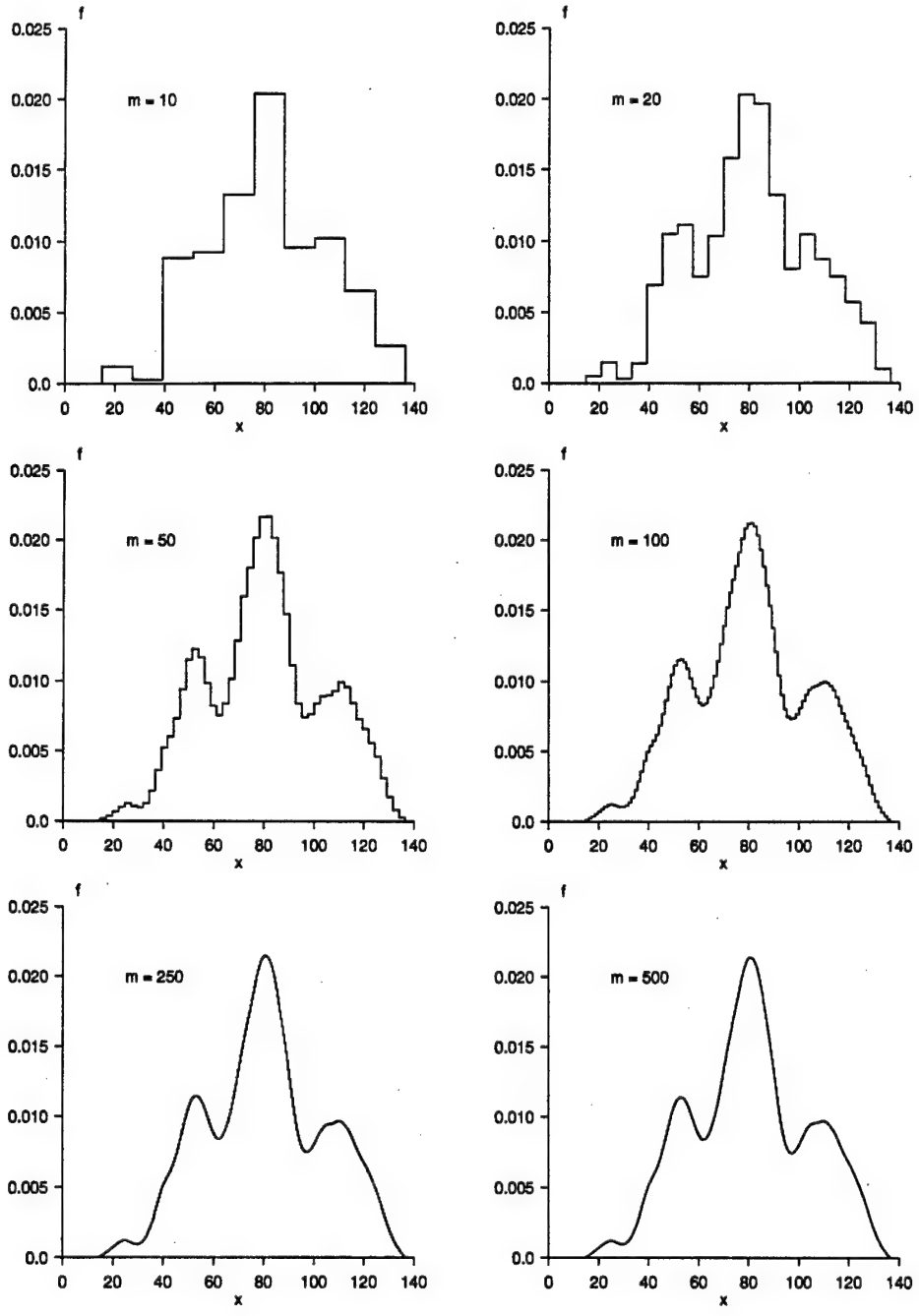

Figure 4.1. AR Density Estimates with Various Discretization Grid Sizes, Buffalo Snowfall Data, $n = 63$.

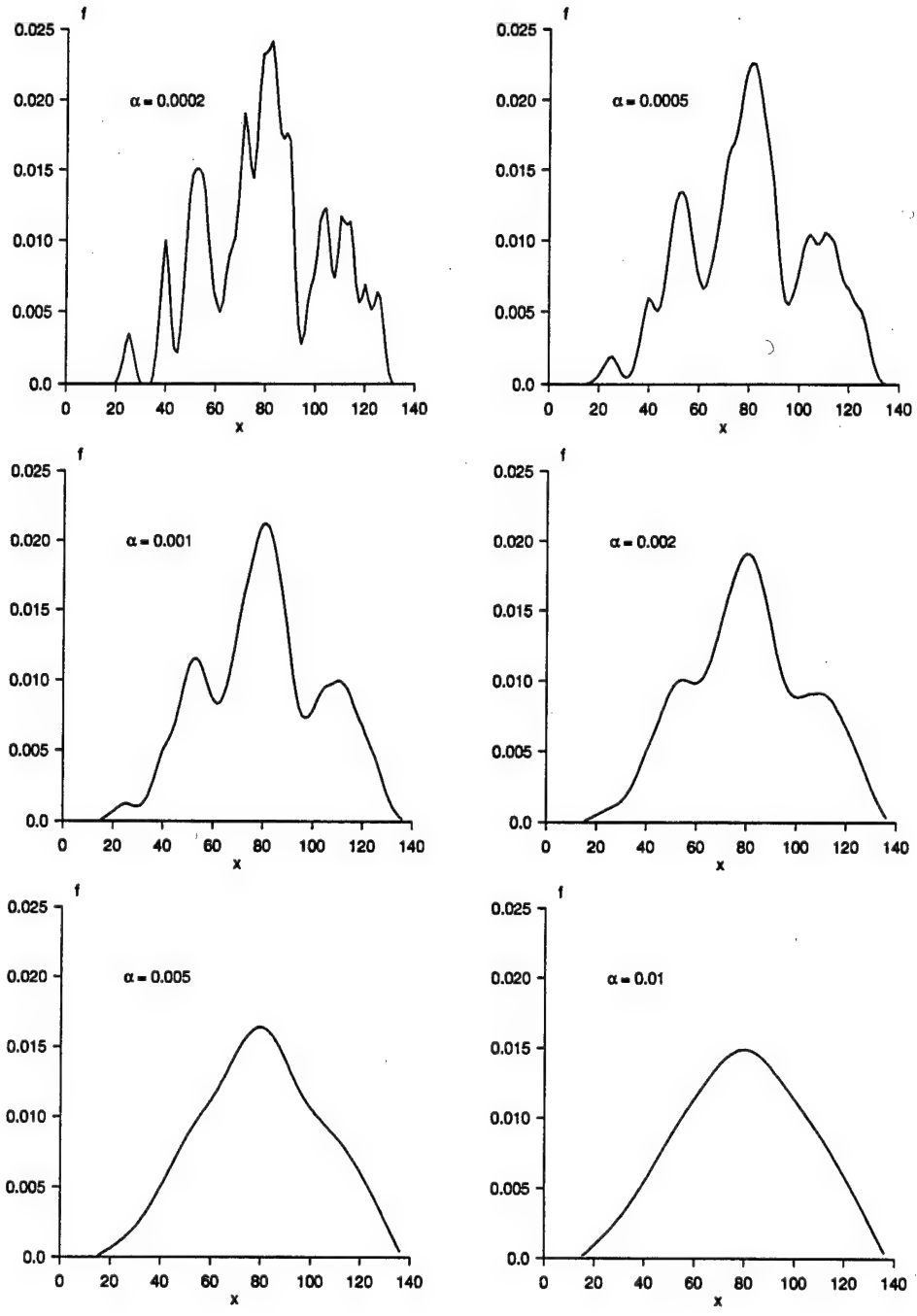

Figure 4.2. AR Density Estimates with Various Smoothing Parameter Values, Buffalo Snowfall Data, $n = 63$.

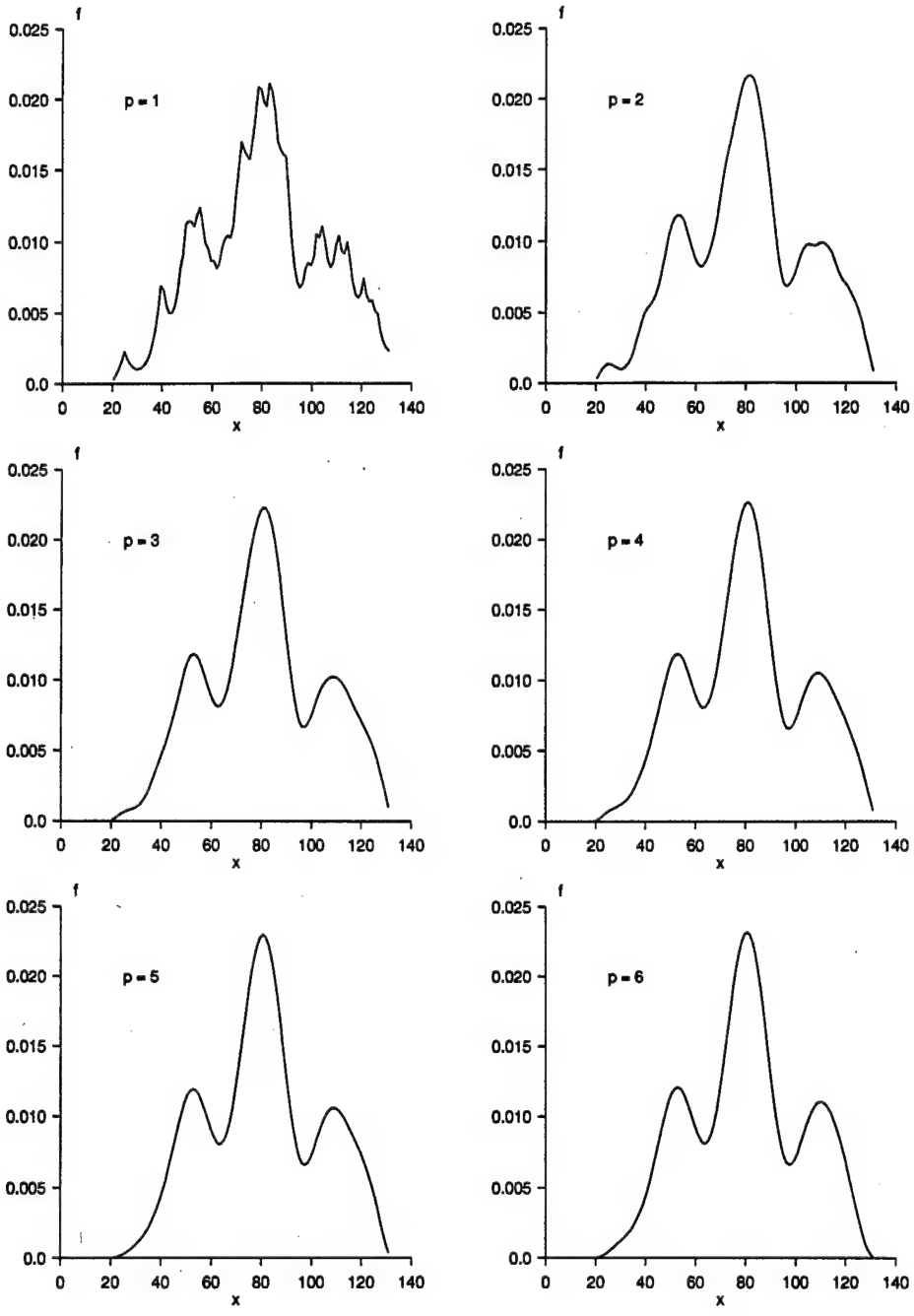

Figure 4.3. AR Density Estimates with Various Penalty Functional Orders, Buffalo Snowfall Data, $n = 63$.

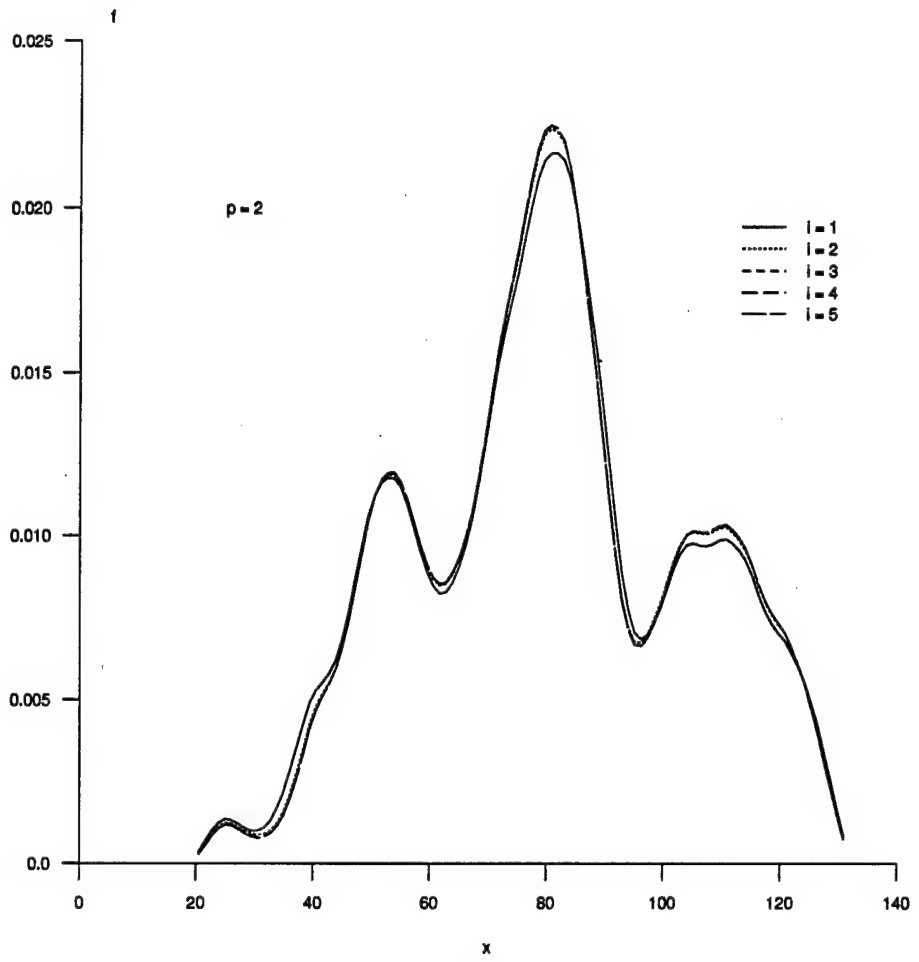

Figure 4.4. Recursive AR Density Estimate Sequence, Buffalo Snowfall Data, $n = 63$.

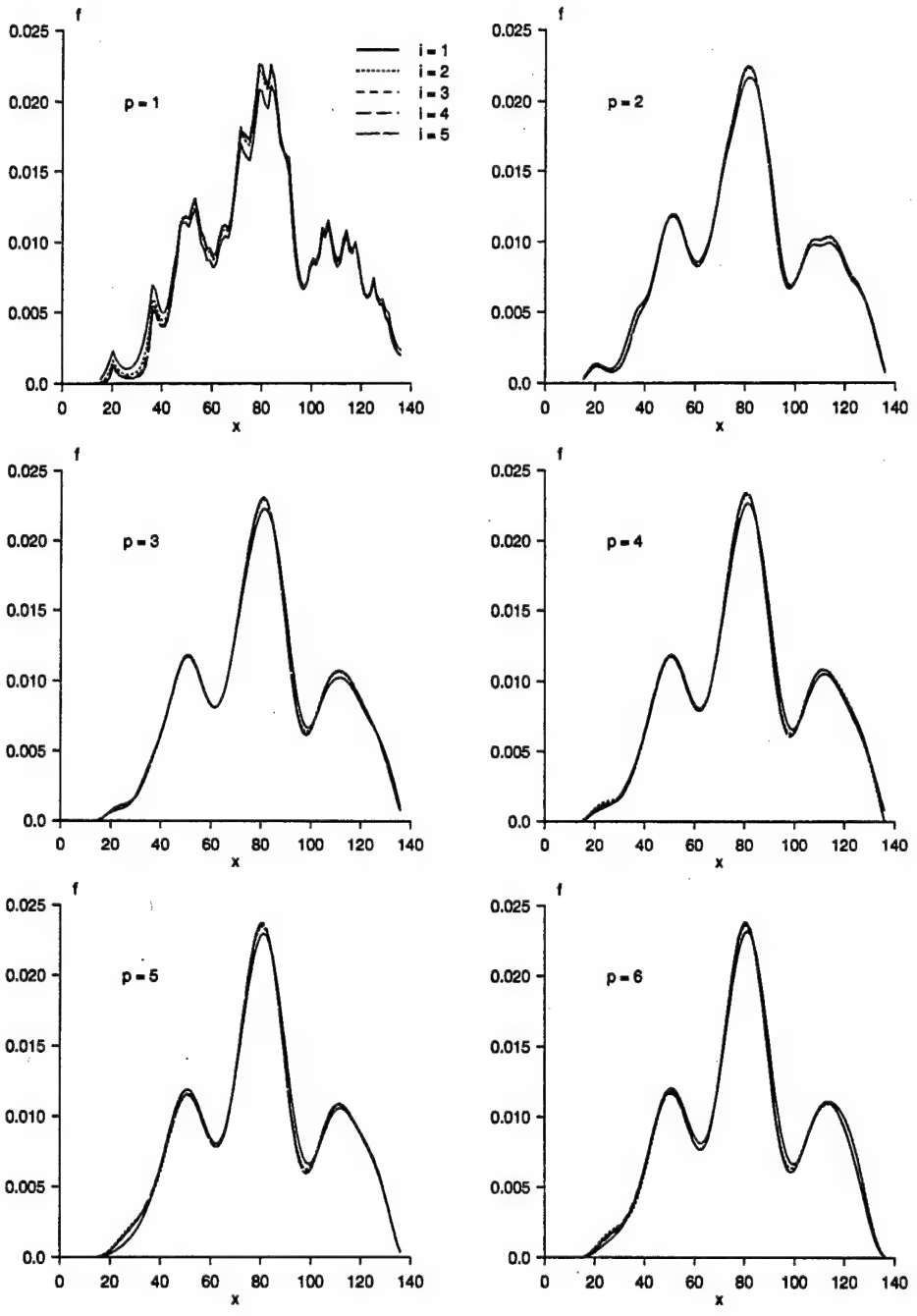

Figure 4.5. Recursive AR Density Estimate Sequences for Various Penalty Functional Orders, Buffalo Snowfall Data, $n = 63$.

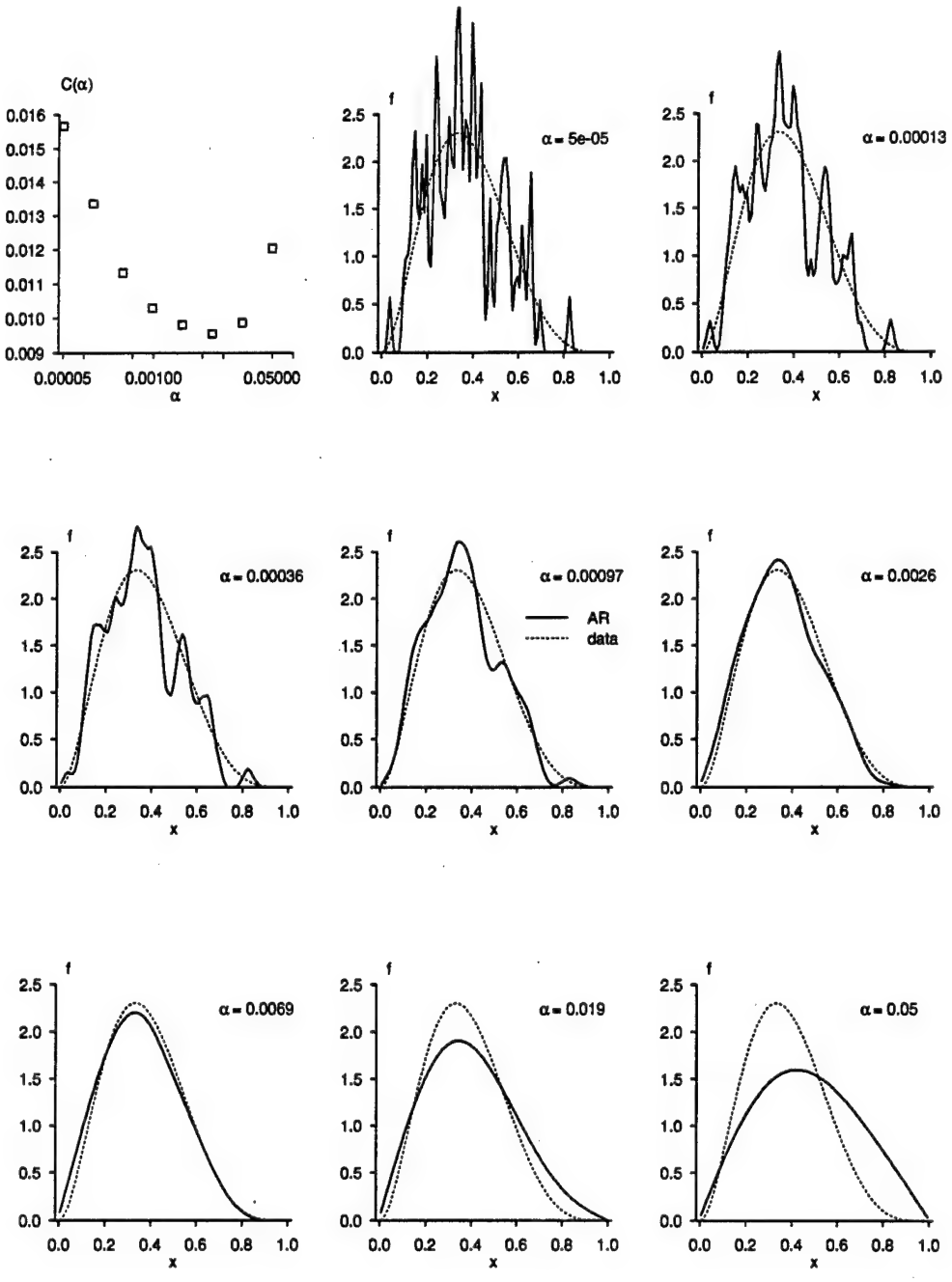

Figure 4.6. GCV Score and AR Density Estimates, $\beta(\cdot, 3, 5)$, $n = 100$.

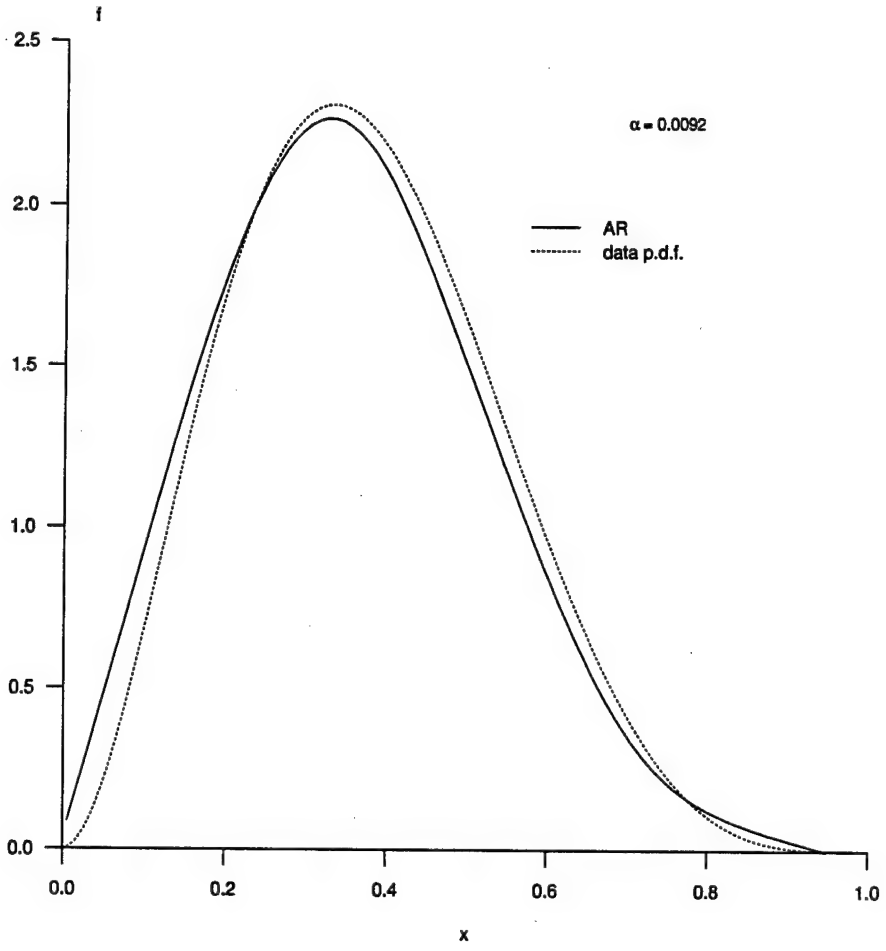

Figure 4.7. AR-GCV Density Estimate, $\beta(\cdot, 3, 5)$, $n = 100$.

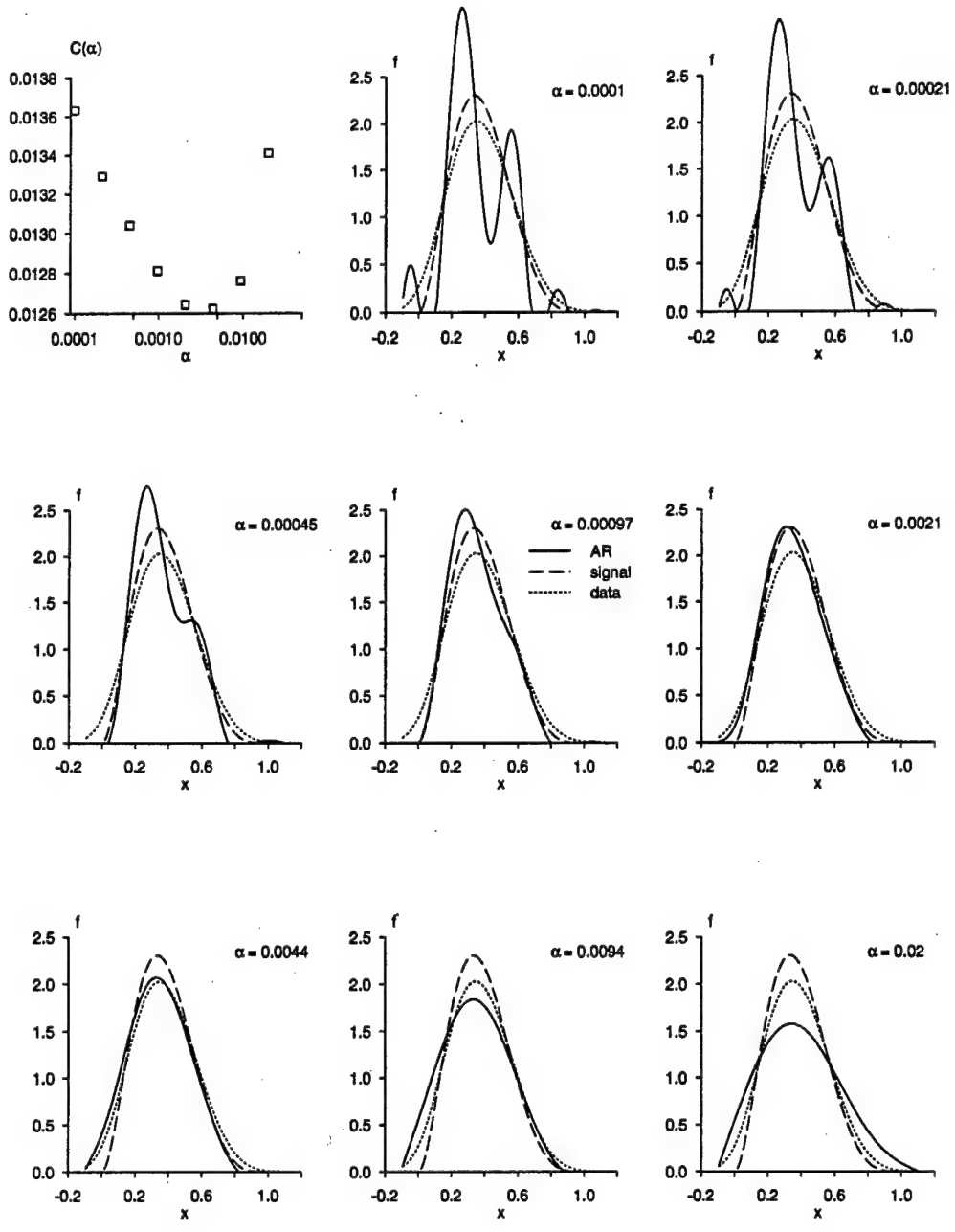

Figure 4.8. GCV Score and AR Deconvolution Estimates, $\beta(\cdot, 3, 5) * \phi(\cdot; 0.1)$, $n = 100$.

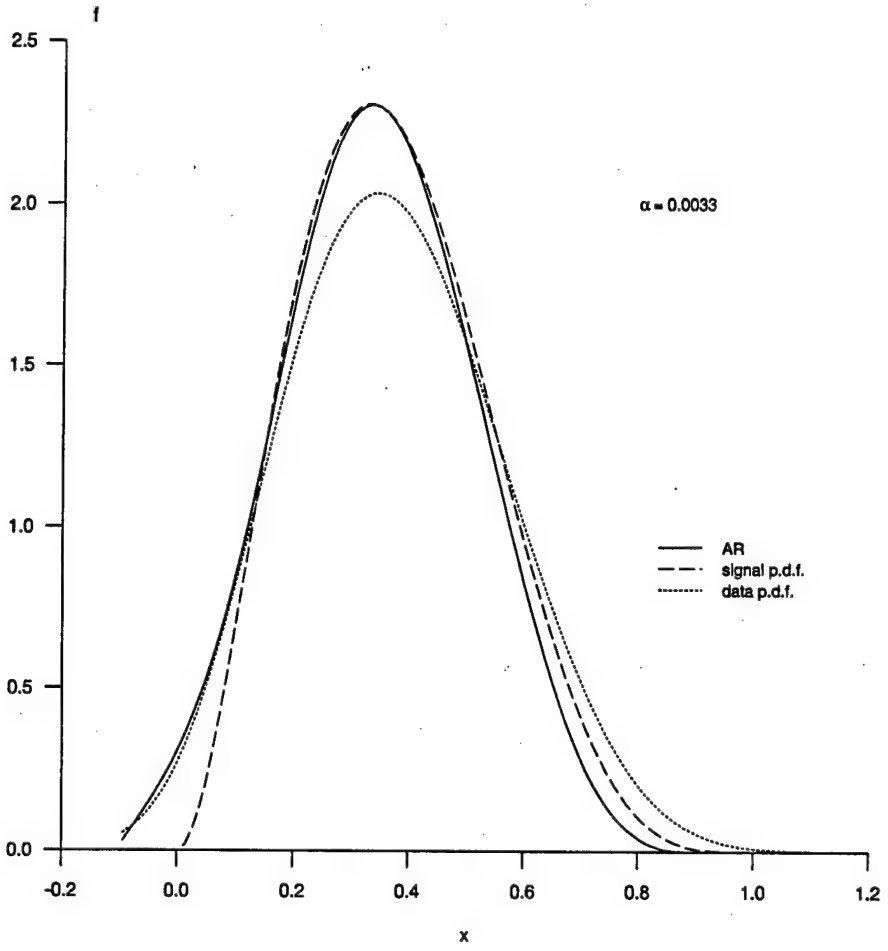

Figure 4.9. AR-GCV Deconvolution Estimate, $\beta(\cdot, 3, 5) * \phi(\cdot; 0.1)$, $n = 100$.

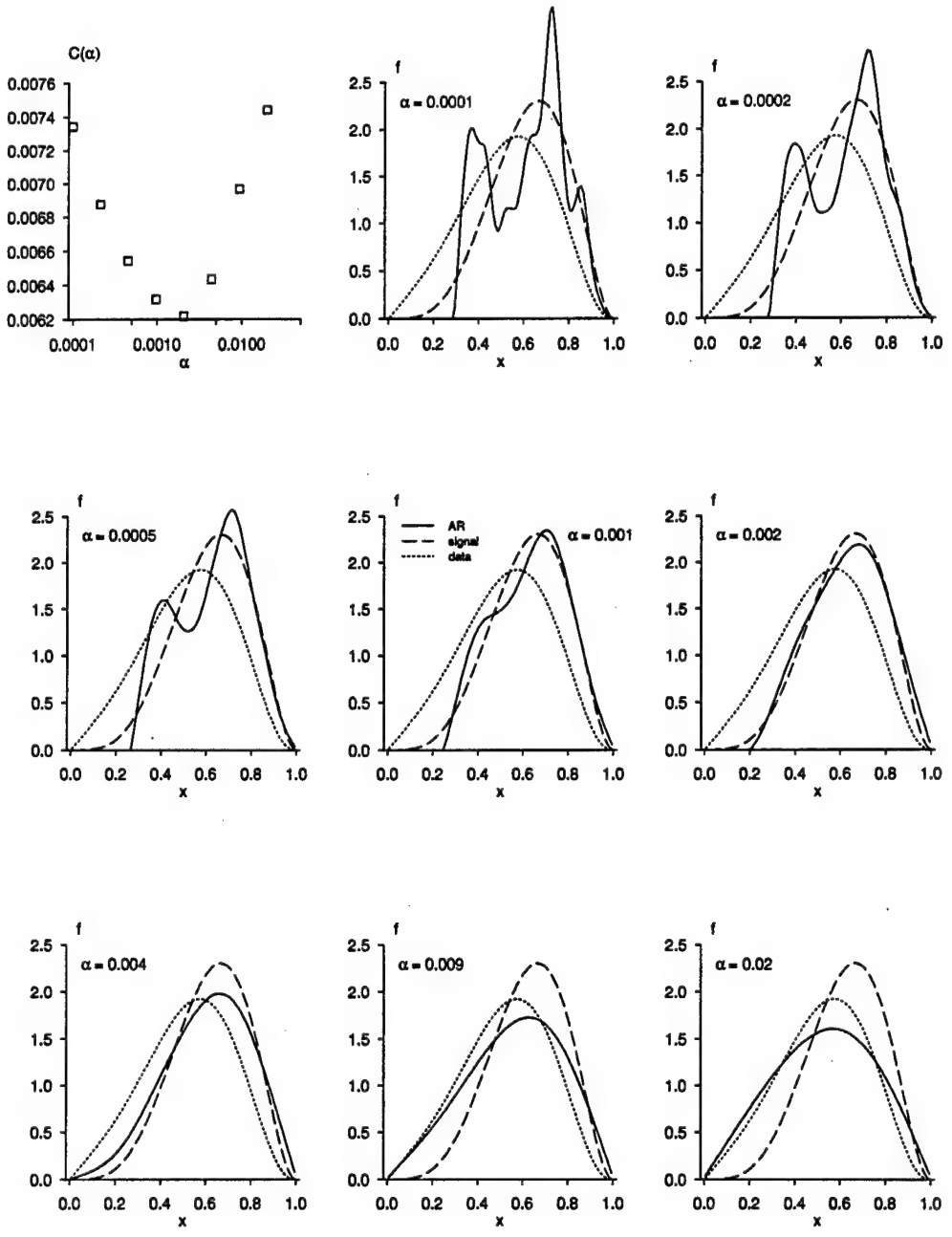

Figure 4.10. GCV Score and AR Corpuscle Estimates, $\beta(\cdot, 5, 3)$, $n = 250$.

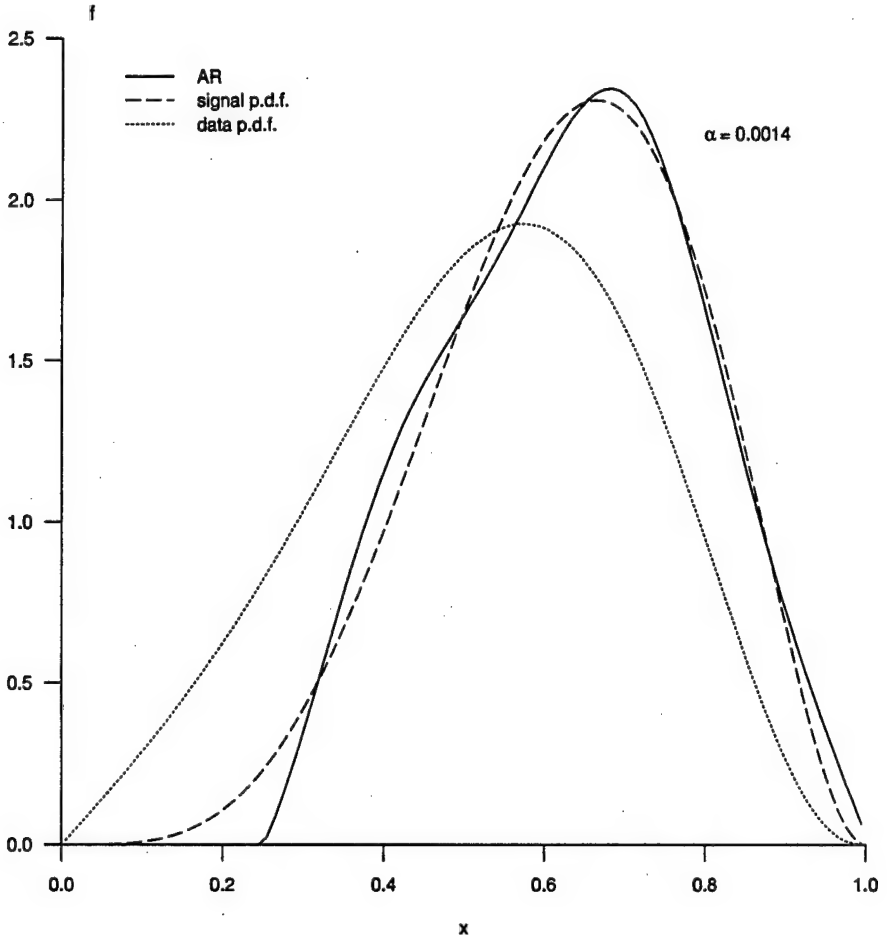

Figure 4.11. AR-GCV Corpuscle Estimate, $\beta(\cdot, 5, 3)$, $n = 250$.

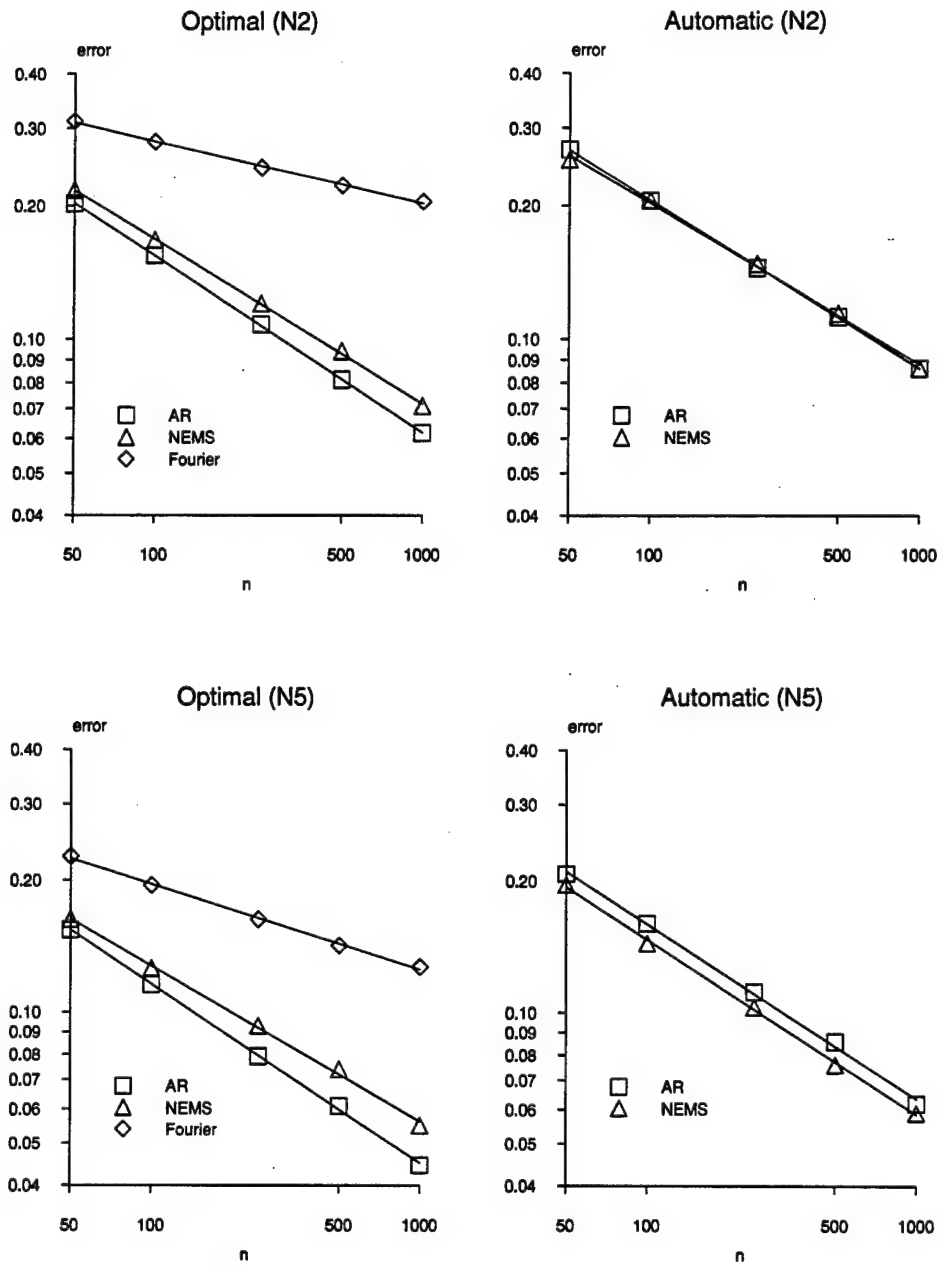

Figure 4.12. Empirical L_1 Error for Fourier, NEMS, and AR Deconvolution Estimators.

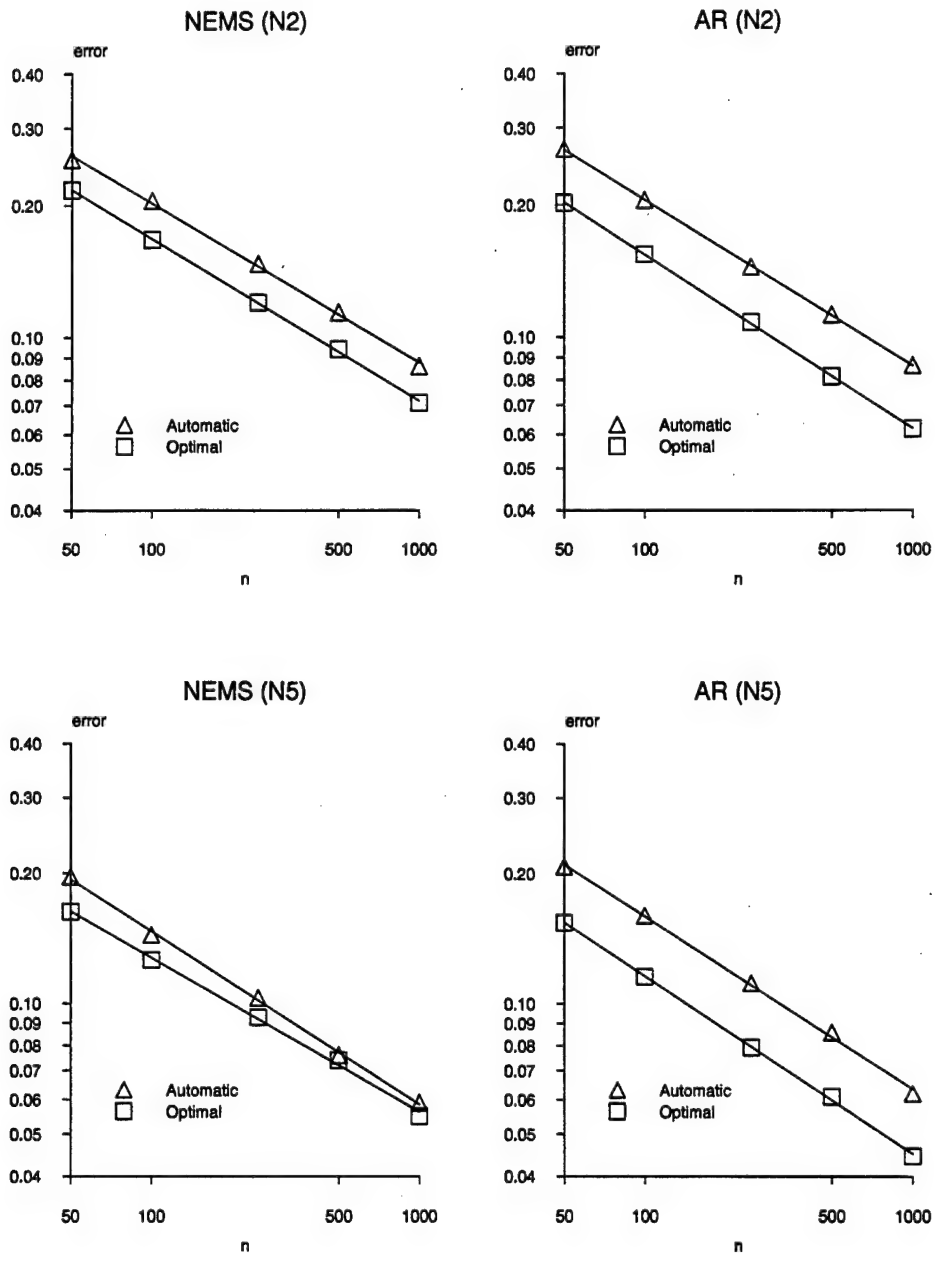

Figure 4.13. Empirical L_1 Error for NEMS and AR Deconvolution Estimators.

Intentionally Left Blank.

Appendix. Estimation for Gaussian Processes

We consider random variables $X : \Omega \to V$, where V is a space of real-valued functions on some set I. Typically $I \subseteq \mathbb{R}$, and here we take $I = [0, 1]$.

Definition (Gaussian Process). The stochastic process X(t) is said to be *Gaussian* if the finite-dimensional distributions of X are multivariate normal; i.e., if for every positive integer n and vector $(t_1, \ldots, t_n) \in I^n$, the vector $(X(t_1), \ldots, X(t_n))$ has a multivariate normal distribution.

A Gaussian process process $X(t) = m(t) + n(t)$ is characterized by its mean value function $\mathrm{E}\, X(t) = m(t)$ and covariance function $K(s,t) = \mathrm{Cov}[X(s), X(t)] = \mathrm{E}\, n(s)\, n(t)$.

Definition (Reproducing Kernel Hilbert Space). A Hilbert space H_K of functions on I with inner product $\langle f, g \rangle_K$ is a *reproducing kernel Hilbert space (RKHS)* if for each $t \in I$ the point evaluation functional V_t, defined by $V_t(f) = f(t)$ for all $f \in H$, is continuous.

In a RKHS, each point evaluation functional V_t has a Riesz representation $V_t(f) = \langle K_t, f \rangle = f(t)$ for a unique $K_t \in H$. The function $K(s,t) \equiv V_t(K_s) = \langle K_t, K_s \rangle_K = K_s(t)$ is called the *reproducing kernel* of H.

The linear span of X is the space

$$L(X) = \left\{ \sum_{i=1}^{n} a_i X(t_i) : n \in \mathbb{N},\ a_i \in \mathbb{R},\ t_i \in I \right\},$$

which is an inner product space with inner product $\langle u, v \rangle = \mathrm{E}(uv)$ and norm $\|u\| = \sqrt{\mathrm{E}\, u^2}$. The Hilbert space generated by $X(t)$, denoted $L_2(X)$, is the $\|\cdot\|$-completion of $L(X)$.

Let H_K be the RKHS with reproducing kernel $K(s,t) = \mathrm{E}[X(s)X(t)]$ and inner product $\langle \cdot, \cdot \rangle_K$. Let $\phi : H_K \to L_2(X)$ be defined by $\phi(K_t) = X(t)$. Then ϕ has the properties

$$\mathrm{E}\, \phi(f) = \langle f, m \rangle_K \quad \text{and} \quad \mathrm{E}\, \phi(f)\phi(g) = \langle f, g \rangle_K .$$

Note that $\{X(t) : t \in I\}$ generates $L_2(X)$, and $\{K_t : t \in I\}$ generates H_K. Furthermore, ϕ is an inner-product-preserving, bijective linear transformation in the sense that

$$\langle \phi(f), \phi(g) \rangle_{L_2(X)} = \langle f, g \rangle_K .$$

Now let $X(t)$ be a stochastic process with mean value function $\mathrm{E}\, X(t) = m(t)$ and known covariance function $K(s,t) = \mathrm{Cov}[X(t), X(s)]$, and let H_K be the RKHS with reproducing kernel K. In this case, the function $\phi : H_K \to L_2(X)$ given by $\phi(K_t) = X(t)$ has the properties

$$\mathrm{E}\, \phi(f) = \langle f, m \rangle_K \quad \text{and} \quad \mathrm{Cov}[\phi(f), \phi(g)] = \langle f, g \rangle_K .$$

Theorem A.1. *Let $\{X(t) : t \in I\}$ be a stochastic process with mean value $m(t) = \mathrm{E}\, X(t)$ and covariance $K(s,t) = \mathrm{Cov}[X(s), X(t)]$, where*

(1) I is countable, or

(2) I is separable, K is continuous, and $X(t)$ is separable.

Let $\mathcal{P}_1$ be the probability measure induced by X on the space of sample paths. Let $\mathcal{P}_0$ be the probability induced by a zero-mean Gaussian process with covariance function K. Then $\mathcal{P}_0$ and $\mathcal{P}_1$ are orthogonal if $m \notin H_K$ and equivalent if $m \in H_K$, in which case

$$\frac{d\mathcal{P}_1}{d\mathcal{P}_0}(X) = \exp\left[\phi(m) - \tfrac{1}{2}\|m\|_K^2\right].$$

Proof. See Parzen [51]. □

The following theorem is useful in that it gives explicit formulas for the RKHS inner products that we need.

Theorem A.2. *Consider the RKHS of functions on $[a,b]$ with $0 \leqslant a < b \leqslant 1$, where the reproducing kernel has the form*

$$K(s,t) = u(s \wedge t)\, v(s \vee t).$$

Let $x(s,t) = u(s\vee t)v(s\wedge t) - u(s\wedge t)v(s\vee t)$, *and let* $y(t) = u'(t)v(t) - u(t)v'(t)$. *If* $x(s,t) > 0$ *and* $y(s) > 0$ *for all* s *and* $t \neq s$ *in* $[a,b]$, *then the corresponding RKHS inner product is given by*

$$\|F\|_*^2 = \int_a^b \frac{[(F/v)']^2}{(u/v)'} + \frac{F(a)^2}{u(a)\,v(a)}.$$

Proof. See Sacks and Ylvisaker [59] for a derivation. We simply verify the reproducing property. Here, $\int$ means $\int ds$.

$$\begin{aligned}
\langle K_t, F\rangle_* &= \int_a^b \frac{(K_t/v)'\,(F/v)'}{(u/v)'} + \frac{K_t(a)\,F(a)}{u(a)\,v(a)} \\
&= \int_a^t \frac{[u\,v(t)/v]'(F/v)'}{(u/v)'} + \int_t^b \frac{[u(t)\,v/v]'(F/v)'}{(u/v)'} + \frac{u(a)\,v(t)\,F(a)}{u(a)\,v(a)} \\
&= v(t)\int_a^t (F/v)' + u(t)\cdot 0 + v(t)F(a)/v(a) \\
&= v(t)[F(t)/v(t) - F(a)/v(a)] + v(t)F(a)/v(a) \\
&= F(t).
\end{aligned}$$

□

Corollary A.3. *Consider the RKHS of functions on* $[a,b]$ *with* $0 \leqslant a < b \leqslant 1$, *where the reproducing kernel has the form*

$$K(s,t) = u(s\wedge t)\,v(s\vee t)\,w(s)\,w(t).$$

The corresponding RKHS inner product is given by

$$\|F\|_+^2 = \|F/w\|_*^2.$$

Proof. Note that $w(s)w(t) = w(s\wedge t)w(s\vee t)$. Apply the theorem to $K(s,t) = u(s\wedge t)w(s\wedge t)\cdot v(s\vee t)w(s\vee t)$; i.e., replace u by uw and v by vw in the form $\|\cdot\|_*^2$ to obtain

$$\|F\|_+^2 = \int_a^b \frac{[(F/(v\,w))']^2}{(u/v)'} + \frac{F(a)^2}{u(a)\,v(a)\,w(a)^2},$$

as required. □

As an aid in performing calculations, we can write

$$\|F\|_*^2 = \int_a^b J + \frac{F(a)^2}{u(a)\,v(a)},$$

where $J = [(F/v)']^2/(u/v)'$. Since $(u/v)' = y/v^2$, the integrand is

$$\begin{aligned} J &= \frac{v^2}{y} \cdot \frac{F'^2 v^2 - 2F'Fv'v + F^2 v'^2}{v^4} \\ &= \frac{1}{y} \cdot \left(F'^2 - v' \cdot \frac{2F'Fv - F^2 v'}{v^2} \right) \\ &= \frac{1}{y} \cdot \left[F'^2 - v' \left(\frac{F^2}{v} \right)' \right]. \end{aligned}$$

We consider several important examples of RKHS's with this type of inner product structure.

Example (1). Let $K(s,t) = G(s \wedge t)$, where G is non-negative and increasing on I. Then $u(t) = G(t)$ and $v(t) = 1$. Observe that $x(s,t) = G(s \vee t) - G(s \wedge t) > 0$ if $s \neq t$, $y(t) = G'(t) > 0$ for all t, and $v' = 0$. So

$$J = \frac{F'^2}{G'},$$

and the quadratic form is

$$\|F\|^2 = \int_a^b \frac{(F')^2}{G'} + \frac{F(a)^2}{G(a)}. \tag{A.1}$$

Example (2). Let $K(s,t) = G(s \wedge t) - G(s)G(t)$, where G is non-negative and increasing on I. Then $u(t) = G(t)$ and $v(t) = 1 - G(t)$. Observe that

$$\begin{aligned} x(s,t) &= G(s \vee t)[1 - G(s \wedge t)] - G(s \wedge t)[1 - G(s \vee t)] \\ &= G(s \vee t) - G(s \wedge t), \end{aligned}$$

so $x > 0$ if $s \neq t$. Furthermore, $y(t) = G'(t)[1 - G(t)] + G(t)G'(t) = G'(t)$. So $y > 0$ for all t, and $v' = -y$. Thus,

$$J = \frac{F'^2}{G'} + \left(\frac{F^2}{1 - G} \right)'.$$

The quadratic form is

$$\|F\|^2 = \int_a^b \frac{(F')^2}{G'} + \frac{F(b)^2}{1-G(b)} + F(a)^2 \left(\frac{1}{G(a)[1-G(a)]} - \frac{1}{1-G(a)} \right)$$

or

$$\|F\|^2 = \int_a^b \frac{(F')^2}{G'} + \frac{F(a)^2}{G(a)} + \frac{F(b)^2}{1-G(b)}. \tag{A.2}$$

Example (3). Let $K(s,t) = G(s)G(t)(s \wedge t - st)$, where G is non-negative and increasing on I. Then $u(t) = t$, $v(t) = 1-t$, $w(t) = G(t)$, and the quadratic form is

$$\|F\|^2 = \int_a^b \left[\left(\frac{F}{G} \right)' \right]^2 + \frac{1}{a} \left[\frac{F(a)}{G(a)} \right]^2 + \frac{1}{1-b} \left[\frac{F(b)}{G(b)} \right]^2 \tag{A.3}$$

or

$$\|G \cdot F\|^2 = \int_a^b (F')^2 + \frac{F(a)^2}{a} + \frac{F(b)^2}{1-b}.$$

INTENTIONALLY LEFT BLANK.

Bibliography

[1] Adams, R. A., *Sobolev Spaces*, Academic Press, Inc., San Diego, 1978.

[2] Aronszajn, N., "Theory of Reproducing Kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[3] Atteia, M., *Hilbertian Kernels and Spline Functions*, Elsevier Science Publishers B. V., Amsterdam, 1992.

[4] Bennett, C. A., *Asymptotic Properties of Ideal Linear Estimators*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1952.

[5] Billingsley, P., *Convergence of Probability Measures*, John Wiley and Sons, Inc., New York, 1968.

[6] Bosq, D. and J.-P. Lecoutre, *Theorie de l'Estimation Fonctionelle*, Economica, Paris, 1987.

[7] Carroll, R. and D. Ruppert, *Transformation and Weighting in Regression*, Chapman and Hall, New York, 1988.

[8] Cox, D., "Approximation of Least Squares Regression on Nested Subspaces," *The Annals of Statistics*, vol. 16 (2), pp. 713–732, 1988.

[9] Cox, D., "Approximation of Method of Regularization Estimators," *The Annals of Statistics*, vol. 16 (2), pp. 694–712, 1988.

[10] Cox, D. and F. O'Sullivan, "Asymptotic Analysis of Penalized Likelihood and Related Problems," *The Annals of Statistics*, vol. 18 (4), pp. 1676–1695, 1990.

[11] Csörgő, M. and P. Révész, *Strong Approximations in Probability and Statistics*, Academic Press, New York, 1981.

[12] David, H. A., *Order Statistics*, John Wiley and Sons, Inc., New York, 1981.

[13] de Montricher, G. F., R. A. Tapia, and J. R. Thompson, "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *The Annals of Statistics*, vol. 3 (6), pp. 1329–1348, 1975.

[14] Devroye, L., "The Equivalence of Weak, Strong, and Complete Convergence in L_1 for Kernel Density Estimates," *The Annals of Statistics*, vol. 11 (3), pp. 896–904, 1983.

[15] Eggermont, P. P. B. and V. N. LaRiccia, "Nonlinearly Smoothed EM Density Estimation with Automated Smoothing Parameter Selection for Nonparametric Deconvolution Problems," *Journal of the American Statistical Association*, vol. 92 (440), pp. 1451–1458, Dec 1997.

[16] Eubank, R. L., *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc., New York, 1988.

[17] Eubank, R. L. and V. N. LaRiccia, "Weighted L^2 Quantile Distance Estimators for Randomly Censored Data," *Journal of Multivariate Analysis*, vol. 14 (3), pp. 621–624, Jun 1984.

[18] Fan, J., "Global Behavior of Deconvolution Kernel Estimates," *Statistica Sinica*, vol. 1, pp. 541–555, 1991.

[19] Fisher, R. A., "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 222, pp. 309–368, 1921.

[20] Gauss, K. F., "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae," *Commentationes Soc. Reg. Gottingensis*, vol. 5, pp. 33–90, 1822.

[21] Good, I. J. and R. A. Gaskins, "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, vol. 58 (2), pp. 255–277, 1971.

[22] Grenander, U., *Abstract Inference*, John Wiley and Sons, New York, 1981.

[23] Gu, C., "Rkpack and Its Applications: Fitting Smoothing Spline Models," Tech. Rep. 587, Department of Statistics, University of Wisconsin, Madison, WI, Jan 1989.

[24] Gu, C., "Adaptive Spline Smoothing in Non-Gaussian Regression Models," *Journal of the American Statistical Association*, vol. 85 (411), pp. 801–807, 1990.

[25] Hájek, J., "Local Asymptotic Minimax and Admissibility in Estimation," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 175–194, University of California Press, Berkeley, CA, 1972.

[26] Hall, P. and R. L. Smith, "The Kernel Method for Unfolding Sphere Size Distributions," *Journal of Computational Physics*, vol. 74, pp. 409–421, 1988.

[27] Härdle, W., *Applied Nonparametric Regression*, Cambridge University Press, 1990.

[28] Inagaki, N., "Asymptotic Relations Between the Likelihood Estimating Function and the Maximum Likelihood Estimator," *Annals of the Institute of Statistical Mathematics*, vol. 25, pp. 1–26, 1973.

[29] Johnstone, I. M. and B. W. Silverman, "Speed of Estimation in Positron Emission Tomography and Related Inverse Problems," *The Annals of Statistics*, vol. 18 (1), pp. 251–280, 1990.

[30] Kallenberg, O., *Random Measures*, Academic Press, Inc., New York, 1983.

[31] Karr, A. F., *Point Processes and Their Statistical Inference*, Marcel Dekker, Inc., New York, 1986.

[32] Kindermann, R. P. and V. N. LaRiccia, "Closed Form Asymptotically Efficient Estimators Based upon Order Statistics," *Statistics and Probability Letters*, vol. 3, pp. 29–34, 1985.

[33] Klonias, V. K., "Consistency of Two Nonparametric Maximum Penalized Likelihood Estimators of the Probability Density Function," *The Annals of Statistics*, vol. 10 (3), pp. 811–824, 1982.

[34] Klonias, V. K., "On a Class of Nonparametric Density and Regression Estimators," *The Annals of Statistics*, vol. 12 (4), pp. 1263–1284, 1984.

[35] Kufner, A., *Weighted Sobolev Spaces*, John Wiley and Sons, Inc., New York, 1985.

[36] Kutoyants, Y. A., *Parameter Estimation for Stochastic Processes*, Helderman Verlag, Berlin, 1984.

[37] Laplace, P. S., "Mémoire sur les Formules Qui Sont Fonctions de Très Grands Nombres et sur Leurs Application aux Probabilités," *Oeuvres de Laplace*, vol. 12, pp. 301–345, 1810.

[38] LaRiccia, V. N., "Asymptotic Properties of Weighted L^2 Quantile Distance Estimators," *The Annals of Statistics*, vol. 10 (2), pp. 621–624, 1984.

[39] LaRiccia, V. N., "Parameter Estimation Based upon Nonparametric Function Estimators," *Communications in Statistics – Theory and Methods*, vol. 13 (22), pp. 2771–2793, 1984.

[40] LaRiccia, V. N. and D. M. Mason, "Optimal Goodness-of-Fit Tests for Location/Scale Families of Distributions Based on the Sum of Squares of L-Statistics," *The Annals of Statistics*, vol. 13 (1), pp. 315–330, 1985.

[41] LaRiccia, V. N. and T. E. Wehrly, "Asymptotic Properties of a Family of Minimum Quantile Distance Estimators," *Journal of the American Statistical Association*, vol. 80 (391), pp. 742–747, 1985.

[42] Le Cam, L., "On the Asymptotic Theory of Estimation and Testing Hypotheses," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 129–156, University of California Press, Berkeley, CA, 1956.

[43] Le Cam, L. and G. Yang, *Asymptotics in Statistics – Some Basic Concepts*, Springer Verlag, New York, 1990.

[44] Lloyd, E. H., "Least-Squares Estimation of Location and Scale Parameters Using Order Statistics," *Biometrika*, vol. 39, pp. 88–95, 1952.

[45] Long, J. C., R. C. Williams, and M. Urbanek, "An EM Algorithm and Testing Strategy for Multiple-Locus Haplotypes," *American Journal of Human Genetics*, vol. 56, pp. 799–810, 1995.

[46] Luenberger, D. G., *Optimization by Vector Space Methods*, John Wiley and Sons, Inc., New York, 1969.

[47] Mendelsohn, J. and J. Rice, "Deconvolution of Microfluorometric Histograms with B Splines," *Journal of the American Statistical Association*, vol. 77 (380), pp. 748–753, 1982.

[48] Nussbaum, M., "Asymptotic Equivalence of Density Estimation and Gaussian White Noise," Tech. Rep., Weierstrass Institute, Berlin, 1996.

[49] Nychka, D., G. Wahba, S. Goldfarb, and T. Pugh, "Cross-Validated Spline Methods for the Estimation of Three-Dimensional Tumor Size Distributions from Observations on Two-Dimensional Cross Sections," *Journal of the American Statistical Association*, vol. 79 (388), pp. 832–846, 1984.

[50] O'Sullivan, F., "A Statistical Perspective on Ill-Posed Inverse Problems," *Statistical Science*, vol. 1 (4), pp. 502–527, 1986.

[51] Parzen, E., "Statistical Inference on Time Series by Hilbert Space Methods, I," Tech. Rep. 23, Applied Mathematics and Statistics Laboratory, Stanford University, Stanford, CA, Jan 1959.

[52] Parzen, E., "An Approach to Time Series Analysis," *Annals of Mathematical Statistics*, vol. 32 (4), pp. 951–989, Dec 1961.

[53] Parzen, E., "Regression Analysis of Continuous Parameter Time Series," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 469–489, University of California Press, Berkeley, CA, 1961.

[54] Parzen, E., "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association*, vol. 74 (365), pp. 105–121, Mar 1979.

[55] Pearson, K., "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London, Series A*, vol. 186, pp. 343–414, 1895.

[56] Pollard, D., *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

[57] Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley and Sons, Inc., New York, 1973.

[58] Riesz, F. and B. Nagy, *Functional Analysis*, Frederick Ungar Publishing Co., New York, 1955.

[59] Sacks, J. and D. Ylvisaker, "Designs for Regression Problems with Correlated Errors," *Annals of Mathematical Statistics*, vol. 37, pp. 66–89, 1966.

[60] Seber, G. A. F., *Linear Regression Analysis*, John Wiley and Sons, Inc., New York, 1977.

[61] Sen, P. K. and J. M. Singer, *Large Sample Methods in Statistics*, Chapman and Hall, New York, 1993.

[62] Serfling, R. J., *Approximation Theorems of Probability and Mathematical Statistics*, John Wiley and Sons, Inc., New York, 1980.

[63] Shorack, G. R. and J. A. Wellner, *Empirical Processes with Applications to Statistics*, John Wiley and Sons, Inc., New York, 1986.

[64] Silverman, B. W., "Density Ratios, Empirical Likelihood and Cot Death," *Applied Statistics*, vol. 27 (1), pp. 26–33, 1978.

[65] Silverman, B. W., "Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives," *The Annals of Statistics*, vol. 6 (1), pp. 177–184, 1978.

[66] Silverman, B. W., "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, vol. 10 (3), pp. 795–810, 1982.

[67] Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1990.

[68] Small, C. G. and D. L. McLeish, *Hilbert Space Methods in Probability and Statistical Inference*, John Wiley and Sons, Inc., New York, 1994.

[69] Stefanski, L. A., "Rates of Convergence of Some Estimators in a Class of Deconvolution Problems," *Statistics and Probability Letters*, vol. 9, pp. 229–235, 1990.

[70] Stone, C. J., "Optimal Rates of Convergence for Nonparametric Estimates," *The Annals of Statistics*, vol. 8 (6), pp. 1348–1360, 1980.

[71] Stone, C. J., "Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics*, vol. 12 (4), pp. 1285–1297, 1984.

[72] Stuart, A. and J. K. Ord, *Kendall's Advanced Theory of Statistics, 5th Edition*, vol. 2, Oxford University Press, New York, 1991.

[73] Thompson, J. R. and R. A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[74] Tikhonov, A. N. and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, V. H. Winston and Sons, Washington, DC, 1977.

[75] van Es, B. and A. Hoogendoorn, "Kernel Estimation in Wicksell's Corpuscle Problem," *Biometrika*, vol. 77 (1), pp. 139–145, 1990.

[76] Wahba, G., "Interpolating Spline Methods for Density Estimation I. Equi-spaced Knots," *The Annals of Statistics*, vol. 3 (1), pp. 30–48, 1975.

[77] Wahba, G., "Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation," *The Annals of Statistics*, vol. 3 (1), pp. 15–29, 1975.

[78] Wahba, G., "Practical Approximate Solutions to Linear Operator Equations When the Data Are Noisy," *SIAM Journal of Numerical Analysis*, vol. 14 (4), pp. 651–667, Sep 1977.

[79] Wahba, G., *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

[80] Wilson, J. D., "A Smoothed EM Algorithm for the Solution of Wicksell's Corpuscle Problem," *Journal of Statistical Computation and Simulation*, vol. 34, pp. 195–221, 1989.

INTENTIONALLY LEFT BLANK.

NO. OF COPIES ORGANIZATION

2 DEFENSE TECHNICAL INFORMATION CENTER
DTIC DDA
8725 JOHN J KINGMAN RD
STE 0944
FT BELVOIR VA 22060-6218

1 DIRECTOR
US ARMY RESEARCH LAB
AMSRL CS AL TP
2800 POWDER MILL RD
ADELPHI MD 20783-1145

1 DIRECTOR
US ARMY RESEARCH LAB
AMSRL CS AL TA
2800 POWDER MILL RD
ADELPHI MD 20783-1145

3 DIRECTOR
US ARMY RESEARCH LAB
AMSRL CI LL
2800 POWDER MILL RD
ADELPHI MD 20783-1145

NO. OF COPIES ORGANIZATION

ABERDEEN PROVING GROUND

4 DIR USARL
AMSRL CI LP (305)

1 DIR USARL
AMSRL SL
DR WADE

1 DIR USARL
AMSRL SL B
LTC GILMAN

20 DIR USARL
AMSRL SL BE
BAKER
BELY
COLLINS (15 CP)
MOSS
SAUCIER
SHNIDMAN

INTENTIONALLY LEFT BLANK.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1998	Final, Jan 96 - Apr 97

4. TITLE AND SUBTITLE

Functional Estimation: The Asymptotic Regression Approach

5. FUNDING NUMBERS

1L162618AH80

6. AUTHOR(S)

Joseph C. Collins III

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
ATTN: AMSRL-SL-BE
Aberdeen Proving Ground, MD 21005-5068

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-1644

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

10.SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Through an appeal to *asymptotic* Gaussian representations of certain empirical stochastic processes, we are able to apply the technique of continuous regression to derive parametric and nonparametric functional estimates for underlying probability laws.

This *asymptotic regression* approach yields estimates for a wide range of statistical problems, including estimation based on the empirical quantile function, Poisson process intensity estimation, parametric and nonparametric density estimation, and estimation for inverse problems.

Consistency and asymptotic distribution theory are established for the general parametric estimator. In the case of nonparametric estimation, we obtain rates of convergence for the density estimator in various norms.

We demonstrate the application of this methodology to inverse problems and compare the performance of the asymptotic regression estimator to other estimation schemes in a simulation study. The asymptotic regression estimates are easily computable and are seen to be competitive with other results in these areas.

14. SUBJECT TERMS

statistics, parametric and nonparametric estimation, density estimation, inverse problems

15. NUMBER OF PAGES

135

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

INTENTIONALLY LEFT BLANK.

USER EVALUATION SHEET/CHANGE OF ADDRESS

This Laboratory undertakes a continuing effort to improve the quality of the reports it publishes. Your comments/answers to the items/questions below will aid us in our efforts.

1. ARL Report Number/Author ARL-TR-1644 (Collins) Date of Report March 1998

2. Date Report Received _____

3. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which the report will be used.) _____

4. Specifically, how is the report being used? (Information source, design data, procedure, source of ideas, etc.) _____

5. Has the information in this report led to any quantitative savings as far as man-hours or dollars saved, operating costs avoided, or efficiencies achieved, etc? If so, please elaborate. _____

6. General Comments. What do you think should be changed to improve future reports? (Indicate changes to organization, technical content, format, etc.) _____

CURRENT ADDRESS

_____
Organization

_____
Name E-mail Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

7. If indicating a Change of Address or Address Correction, please provide the Current or Correct address above and the Old or Incorrect address below.

OLD ADDRESS

_____
Organization

_____
Name

_____
Street or P.O. Box No.

_____
City, State, Zip Code

(Remove this sheet, fold as indicated, tape closed, and mail.)
(DO NOT STAPLE)

DEPARTMENT OF THE ARMY

OFFICIAL BUSINESS

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO 0001,APG,MD

POSTAGE WILL BE PAID BY ADDRESSEE

DIRECTOR
US ARMY RESEARCH LABORATORY
ATTN AMSRL WM BE
ABERDEEN PROVING GROUND MD 21005-5066